Estimating survival in left filtered data

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 were themselves subject to the confounding structures they sought to characterize.

If conditional sequential ignorability of exposure and censoring status at each point in follow-up and positivity can be attained given covariate history, 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 the incomplete observation of cancer incidence outcomes at every point of follow-up over the study period. Nonetheless, we wish to make inferences about the carcinogenicity of MWF exposure over an individual's lifetime starting upon entry into the workforce. The UAW-GM Cohort included those hired roughly between 1938 and 1985. However, cancer incidence reporting at the Michigan Cancer Registry did not begin until

1985. Hence, our observed cohort data is subject to *left filtering*: cancer incidence is the outcome of interest, but 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 was 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 using left filtered data in the presence of 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 survival curves under the following interventions were computed or estimated in 250 replicates: (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. Estimator bias was evaluated by comparing the average of 250 estimates of to the average difference in cancer-free survival under the two rules to the true average survival difference when exposure and censoring were controlled deterministically.

The present project replicates the simulation and bias analyses presented in Chapter 3 of Izano (2017), embeds the problem in the non-parametric structural causal approach of Pearl (1995), and applies the WKM and AWKM estimators to the estimation of colon cancer survival under exposure/censoring rules in the UAW-GM Cohort.

Methods

Causal model

The UAW-GM Cohort data included person-year level exposure, outcome, and covariate data starting at hire. To emulate the shape of the data for this longitudinal cohort, we considered 20 years of data over time indexed by years since hire. Notation representing the variables of interest are presented in Table 1. The causal model represents hypothetical relationships between variables over time compatible with the theory underlying the HWSE in longitudinal occupational cohort studies. At each time point, the effect of cumulative exposure $\bar{A}(t)$ on cancer incidence $Y^(t)$ is confounded by a path through employment status $N(t)$ and underlying health $H(t)$ as well as the path through past exposure $\bar{A}(t-1)$ and vital status $D(t)$. These paths follow straightforward logic: occupational exposure depends upon employment status and past exposure; mortality status is affected by past exposure and cancer history. Confounding by baseline covariates W is assumed throughout.

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

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 $\bar{X}_i(t) = (X_i(k))_{k=1}^t$. 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; susceptibility S and underlying health status H are not known:

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

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.

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

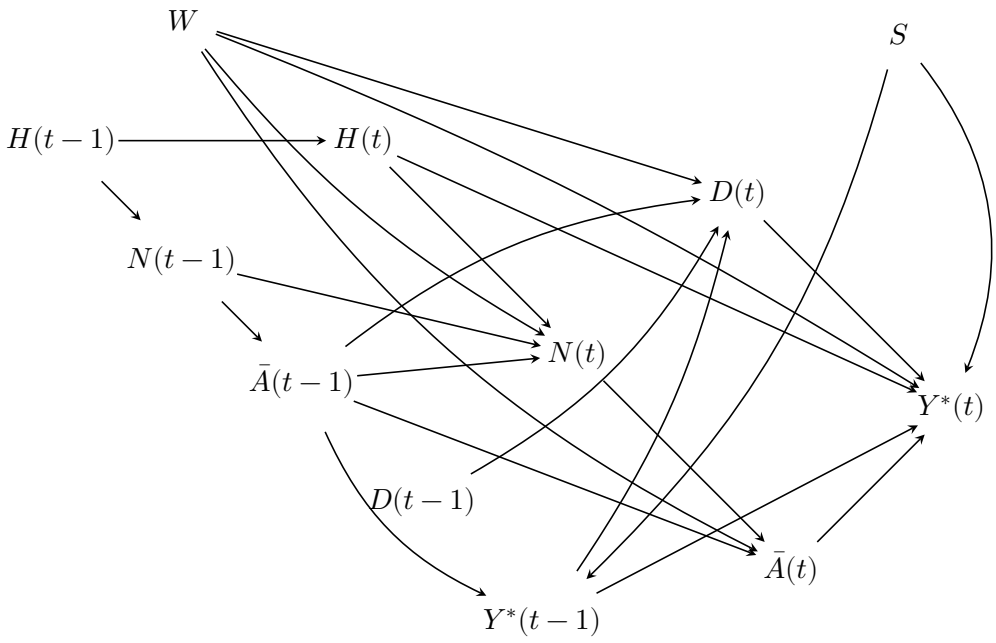


Figure 1 clarifies our conceptualization of the HWSE. At each time point t, the cumulative effect of exposure $\bar{A}(t)$ on cancer incidence $Y^*(t)$ is confounded by employment status $N(t)$ through the backdoor path $\bar{A}(t) \leftarrow N(t) \leftarrow H(t) \rightarrow Y^*(t)$. In the absence of observed data on health status $H(t)$, an analyst may be tempted to condition on employment status $N(t)$, but doing so would introduce collider bias while blocking the causal path between past exposure $\bar{A}(t-1)$ and cancer $Y^*(t)$. Furthermore, an analysis starting at an arbitrary time point after the time origin (hire) would be tantamount to conditioning on those still alive at that time, which would also result in both collider bias and the conditioning on nodes on the causal path between the exposure and outcome of interest.

Simulation

To generate data compatible with our structural causal model, we imposed parametric relationships between the variables. For the $n = 50\,000$ units over $T = 20$ years, we have:

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

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

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

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

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

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

- 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 $\beta_{\bar{Y}}^D$. 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.

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
β_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

Interventions, potential outcomes, target parameters, and estimation

The substantive question of interest was the causal effect of occupational exposure to MWF on cancer incidence risk. Since occupational MWF exposure occurs only when individuals are at work, we defined 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 prevented censoring by death as if it were intervenable. The causal effect was 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 was 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 $\mu_a = \sum_0^K S_a(t)\, dt$. The parameter used in the estimation of bias was the 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 was evaluated by comparing estimates of ψ to its true value in 250 simulations per scenario (the original analysis performed 500). The true value was

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 were 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

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

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:

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

The WKM survival estimator for rule a is

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

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:

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

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

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

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:

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

for all times $t' \geq t$, 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.

Estimation of weights

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

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

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

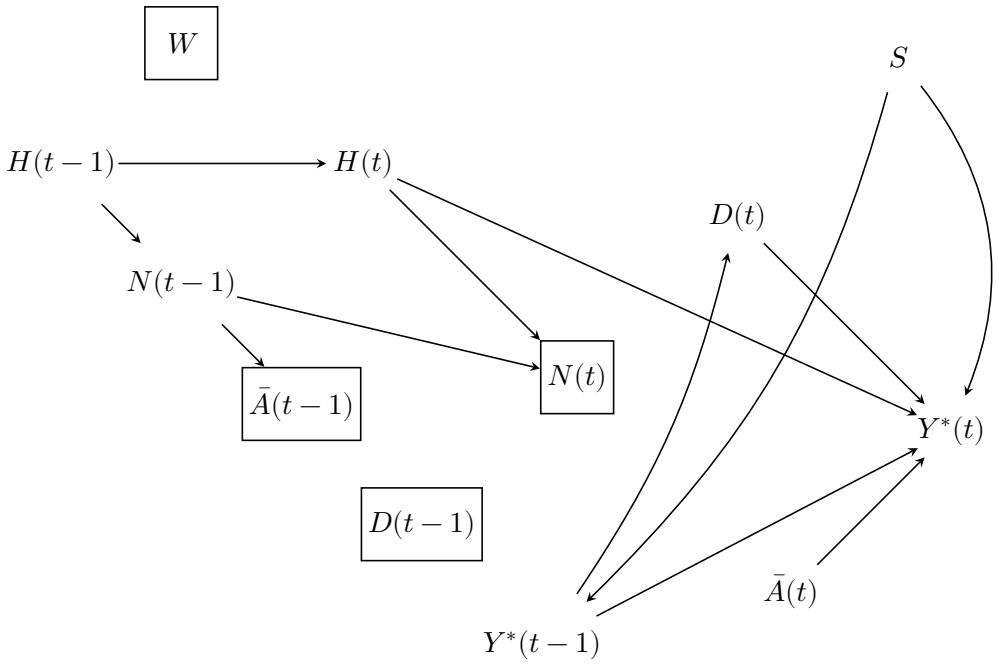
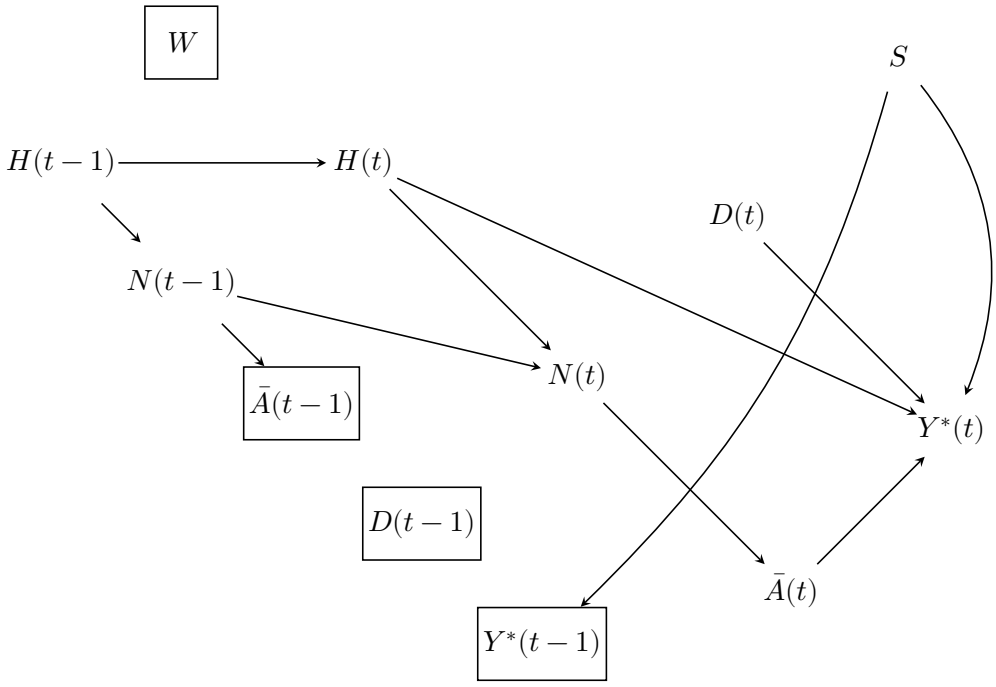


Figure 3: Directed acyclic graph representing the causal relationships encoded in the non-parametric structural equation model at time t after conditioning on $\{W,\ \bar{A}(t-1),\ D(t-1),\ Y^*(t-1)\}$.



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

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

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-estimated survival whereas the the AWKM survival curve was much closer to the truth. The bias of the AWKM survival estimator appeared to be larger in earlier follow-up and smaller for later follow-up time points. The bias of both the the WKM and the AWKM estimators appeared 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 were consistent with the qualitative interpretations of Figure 4. The WKM estimator over-estimated 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 of the WKM-derived contrast was several times larger in magnitude than that of the AWKM-derived contrast. The bias of both estimators was greatest for Scenario 5.

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

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.

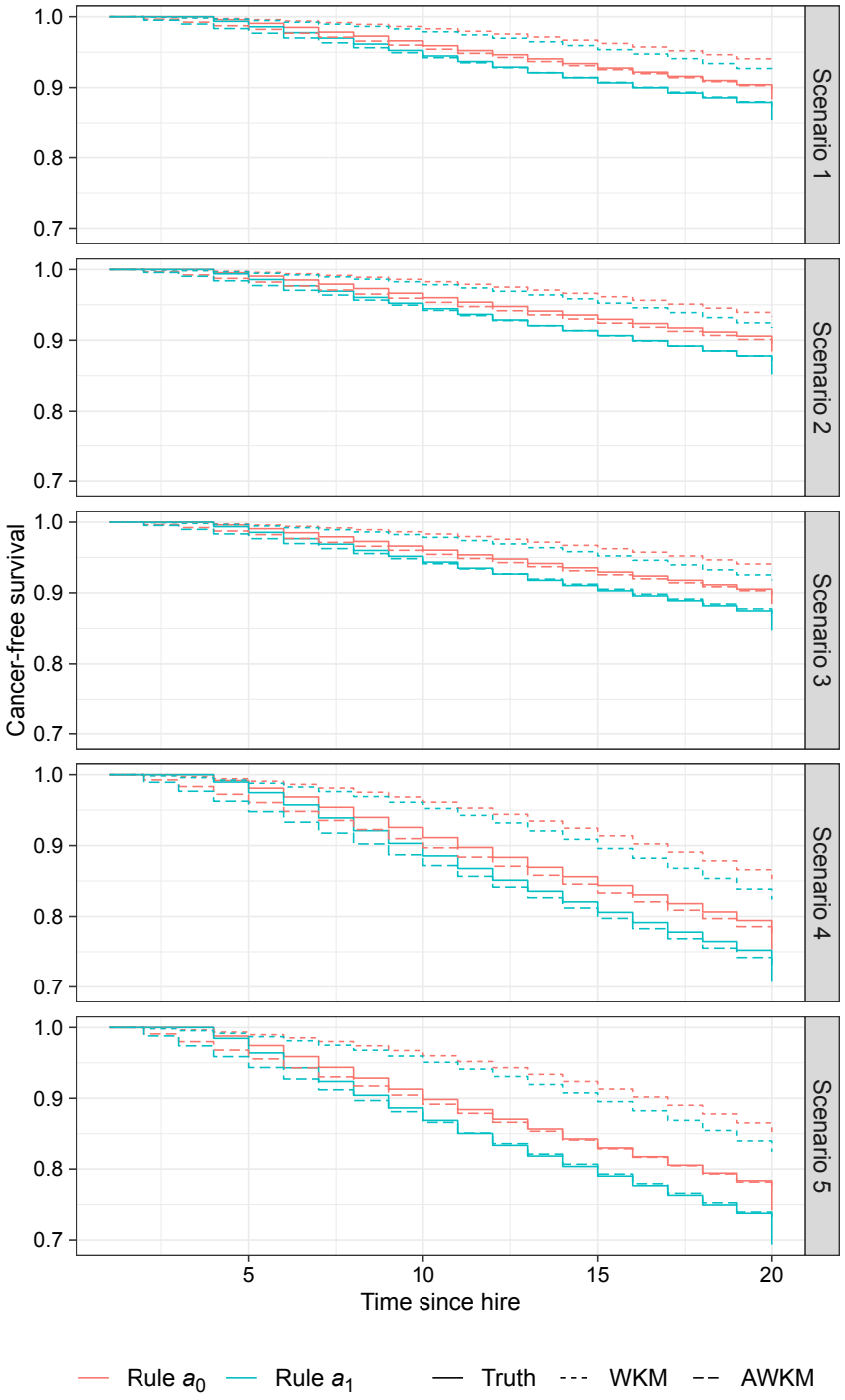


Table 3: True cancer-free survival time μ_a over 20-year follow-up and estimator averages over 250 replicates.

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	19.52	18.91	18.90
	AWKM	19.01	19.00	19.01	17.80	17.71
a_1	Truth	18.80	18.79	18.76	17.54	17.29
	WKM	19.41	19.39	19.39	18.68	18.67
	AWKM	18.76	18.76	18.74	17.30	17.22

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	-0.11	-0.12	-0.13	-0.23	-0.23
AWKM	-0.25	-0.24	-0.27	-0.50	-0.50

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	0.17	0.19	0.21	0.25	0.31
AWKM	0.03	0.07	0.07	-0.02	0.04

Application to the UAW-GM Cohort

In the simulation study, we showed that under several scenarios compatible with our hypothetical causal structure, the AWKM survival estimator had small bias compared to the WKM estimator. The bias was smaller when the cumulative incidence of the outcome was low and at later follow-up time points. Next, we estimated cancer-free survival in a real-world context. Using data from the UAW-GM Cohort study, we followed 38 553 individuals starting from hire to 40 years after hire for incidence of colon cancer. This duration of follow-up spans the vast majority of individuals' working lifetimes. As in the simulation, the UAW-GM data were longitudinal data with baseline covariates, time-varying covariates, and a survival outcome.

The exposures of interest were MWF of three types: straight, soluble, and synthetic (Byers 2006; F. Mirer 2003; F. E. Mirer 2010). Straight MWFs are hydrocarbon-based fluids whose use became widespread by the 1920s. They continue to be widely used because of their simple formulation. In straight MWFs, hydrocarbons of different lengths are mixed together with other additives to attain different properties. Straight MWFs contain polycyclic aromatic hydrocarbons, long known to be carcinogenic (IARC 1973). Soluble oils are water-based oil emulsions first introduced in response to rising oil prices. They now make up the largest market share of MWFs (Childers 2006). Soluble oils are vulnerable to microbial contamination, so the addition of biocides is needed. However, their high lubricity makes soluble MWFs the most popular fluid type. Synthetic MWFs have the best toxicological profile, have no oil, and have a higher resistance to microbial growth. They were introduced into the MWF market in the second half of the 20th C., but fail to out-perform soluble MWFs in metalworking applications. Soluble and synthetic MWFs contain biocides, corrosion inhibitors, and chlorinated compounds, some classified as carcinogenic by the IARC (IARC 1987).

The outcome of interest was colon cancer incidence. There is little past research linking colon cancer to MWF exposures, but there is some evidence suggesting that straight MWFs cause colon cancer (Izano et al. 2019). We obtained cancer incidence data via linkage to the Michigan Cancer Registry, which recorded cancer incidence cases starting on January 1, 1985. The cohort is comprised of individuals hired between 1938 and 1985. Cancer-free survival to the start of the registry was a left-filtering process possibly in the presence of the HWSE as investigated in the simulations. Over the 40 year follow-up period, vital status was obtained through the Social Security Administration, the National Death Index, as well as records provided by the UAW. Baseline confounders included race, sex, plant, and year of hire. Time-varying confounders included age, cumulative time off, and exposure to the metalworking fluids not under study. These terms were included in the estimation of both the treatment and censoring mechanisms, which were estimated with stratification on every year of follow-up. The exposure rules of interest were identical to those in the simulation study: a_0 never exposed and a_1 always exposed while at work. We assumed that the effect of exposure on health status, mortality, cancer incidence, etc occurs after a lag of 15 years. Similarly, we assumed that the effects of cumulative time off occur after a 15 year lag. Counterfactual survival under rules a_0 and a_1 were estimated using the WKM and AWKM estimators. Summary statistics for the full study population and the colon cancer cases are presented in Table 6.

Assumptions

Since we are working with observational data, the evaluation of the no-interference, causal consistency, ignorability, and positivity assumptions are critical for causal inference. The stability of our estimation depends on positivity, which we assessed qualitatively by examining the distribution of the weights. The no-interference assumption may be problematized by the fact that there were a finite number of job types in the factory setting. If one worker operates a particular metalworking machine, then the the other workers are not able to operate that machine at that time. Instead, they may be assigned to assembly tasks, which have lower MWF exposure opportunities. That said,

Table 6: Study population characteristics.

	Full cohort		Colon cancer cases	
n (person-years)	38 553	(1 402 372)	267	(9 635)
Race (%)				
Black	7 133	(18.5)	83	(31.1)
White	31 420	(81.5)	184	(68.9)
Sex (%)				
Female	4 757	(12.3)	49	(18.4)
Male	33 796	(87.7)	218	(81.6)
Plant (%)				
Plant 1	9 092	(23.6)	66	(24.7)
Plant 2	17 090	(44.3)	107	(40.1)
Plant 3	12 371	(32.1)	94	(35.2)
Ever exposed to MWF (%)	22193	(57.6)	104	(39.0)
Year of hire (mean (SD))	1963	(11.62)	1968	(8.79)
Age (mean (SD))	63.38	(9.08)	67.78	(9.27)
Cumulative time off (mean (SD))	0.04	(0.14)	0.05	(0.15)

since these factories were quite large, there may be approximate independence. The consistency assumption is also problematic. The MWFs of interest are complex chemical mixtures whose composition changes by design and by nature. Over the last several decades, the formulation of MWFs has changed significantly in reaction to performance needs and toxicity concerns (F. Mirer 2003; Byers 2006). The composition of MWFs also undergoes unintentional changes because of the nature of their use: MWFs are often applied in contexts where contamination by other substances and microbes is possible and chemical changes due to heat and pressure are likely. Indeed, there are substantial concerns over the carcinogenicity of the chemical species formed in the MWF mixtures that were not originally added (Hidajat et al. 2020). Concerns regarding the consistency assumption may be abrogated in part by adequate adjustment for secular and factory-level characteristics.

Another key assumption meriting discussion is that of conditional ignorability. In order to achieve identification, even in the absence of left filtering, we need to have conditionally ignorable future exposure status and ignorable future censoring status at each time point given past data. In occupational cohorts, employment status and health history are strong predictors of future death (Häfner 1987; Halliday 2014; Laliotis and Stavropoulou 2018). Logically, major causes of death first act through employment status before they precipitate death. This dynamic is actually a key component in the setup for HWSE. We are therefore relatively confident that conditional ignorability of censoring due to death is attained given covariate, exposure, and cancer history. Our confidence in the conditional ignorability of exposure given history was not as strong. In particular, workers may be assigned to certain tasks based on their specific skills and knowledge, which may be associated with structural privileges that confer a lower risk of deleterious health outcomes. The potential

magnitude of this uncontrolled confounding is likely to be limited, however. The education level of the workers in the cohort was homogeneous, and all cohort members were members of the UAW union, which had uniform procedures in place for equitable access to training, wages, and career advancement (Harbison 1950; Barnard 2005). The presence of UAW policies support the assertion that given time since hire, job types (and therefore exposures) were randomly allocated.

Results

- Foo

Discussion

- FU in GM is much longer than 20 years
- Exclusion restriction on H so that its parents do not include A is not necessary for identification in full data
- Exclusion restriction on D so that its parents do not include H not needed if H is observed and conditioned upon
- Not clear if these were for convenience or what
- What if we increase FU? Lags?
- Greater fineness in exposure definition? Different interventions?
- No way to verify to what extent the simulation informs the real data generating system
- Model specification, double robustness

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.

Barnard, John. 2005. *American Vanguard: The United Auto Workers During the Reuther Years, 1935-1970*. Wayne State University Press.

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

Childers, Jean. 2006. "The Chemistry of Metalworking Fluids." In *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.

Halliday, Timothy J. 2014. "Unemployment and Mortality: Evidence from the PSID." *Soc Sci Med* 113 (July): 15–22. https://doi.org/10.1016/j.socscimed.2014.04.038.

Harbison, Frederick H. 1950. "The General Motors-United Auto Workers Agreement of 1950." *Journal of Political Economy* 58 (5): 397–411.

Häfner, H. 1987. "Unemployment and Health." *Dtsch Med Wochenschr* 112 (37): 1428–32. https://doi.org/10.1055/s-2008-1068265.

Hidajat, Mira, Damien Martin McElvenny, Peter Ritchie, Andrew Darnton, William Mueller, Raymond M Agius, John W Cherrie, and Frank de Vocht. 2020. "Lifetime Cumulative Exposure to Rubber Dust, Fumes and n-Nitrosamines and Non-Cancer Mortality: A 49-Year Follow-up of UK Rubber Factory Workers." *Occup Environ Med* 77 (5): 316–23. https://doi.org/10.1136/oemed-2019-106269.

IARC. 1973. *IARC Monographs on the Evaluation of Carcinogenic Risk of the Chemical to Man: Certain Polycyclic Aromatic Hydrocarbons and Heterocyclic Compounds*. Vol. 3. World Health Organization International Agency for Research on Cancer.

———. 1987. *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans. Overall Evaluations of Carcinogenicity: An Updating of IARC Monographs, Volumes 1 to 42. World Health Organization*. Vol. 7. World Health Organization International Agency for Research on Cancer.

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.

Laliotis, Ioannis, and Charitini Stavropoulou. 2018. "Crises and Mortality: Does the Level of Unemployment Matter?" *Soc Sci Med* 214 (October): 99–109. https://doi.org/10.1016/j.socscimed.2018.08.016.

Mirer, Franklin. 2003. "Updated Epidemiology of Workers Exposed to Metalworking Fluids Provides Sufficient Evidence for Carcinogenicity." *Applied Occupational and Environmental Hygiene* 18 (11): 902–12.

Mirer, Franklin E. 2010. "New Evidence on the Health Hazards and Control of Metalworking Fluids Since Completion of the OSHA Advisory Committee Report." *American Journal of Industrial Medicine* 53 (8): 792–801.

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.