Title
Estimating Causal Effects of Occupational Exposures

Author
Izano, Monika A

Publication Date
2017

Estimating Causal Effects of Occupational Exposures

by

Monika A. Izano

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Epidemiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Ellen A. Eisen, Chair
Professor Alan E. Hubbard
Assistant Professor Patrick T. Bradshaw

Spring 2017

Estimating Causal Effects of Occupational Exposures

Abstract

Estimating Causal Effects of Occupational Exposures

by

Monika A. Izano

Doctor of Philosophy in Epidemiology

University of California, Berkeley

Professor Ellen A. Eisen, Chair

Estimates of the risk of occupational exposures are typically based on observational workplace studies that are subject to bias due to the healthy worker survivor effect (HWSE), a ubiquitous process that results in the healthiest workers accruing the most exposure. This body of work is concerned with the estimation of causal effects of occupational exposures from observational workplace studies, in the context of the HWSE.

We estimate the effect of cumulative exposure to straight, soluble, and synthetic metalworking fluids (MWFs) on the incidence of colon cancer in the United Autoworkers-General Motors (UAW-GM) cohort. We use longitudinal targeted minimum-loss based estimation (TMLE) to compute the 25-year risk difference if always exposed above compared to if always exposed below an exposure cutoff while at work. Exposure cutoffs were selected *a priori* at the median of exposed person-years among colon cancer cases. Risk differences are 0.038 (95% CI = 0.022 to 0.054), 0.002 (95% CI = -0.016 to 0.019), and 0.008 (95% CI = 0.002 to 0.014) for straight, soluble, and synthetic MWFs, respectively. By control of the time-varying confounding on the casual pathway that characterizes the HWSE, TMLE estimated effects that were undetectable in earlier reports.

Most workers in UAW-GM were hired decades before the reporting of incident cancers began. Incident cancers that occurred before the start of reporting were *left filtered*. We show that if ignored, left filtering can lead to downward bias in exposure effect estimates. Further, we propose a novel delayed-entry adjusted Kaplan-Meier estimator that controls for time-varying confounding, and permits delayed risk-set entry. The estimator results in little bias in simulated datasets when the outcome is sufficiently rare.

In addition to *dynamic* (realistic) interventions that assign exposure according to workers' employment status, causal contrasts can be defined under *static* (etiologic) interventions that additionally prevent leaving work. Causal effect estimates of the two classes of interventions can differ substantially. While ideally the choice of intervention would be driven by the research question, in practice it may be dictated by the available data. Furthermore, when estimates of the long-term etiologic effects of occupational exposures are not available, guidelines for exposure limits may be based on studies that estimated effects of realistic interventions. In a simulation study we investigate the conditions under which the two effect measures are comparable, and identify factors that drive the differences between the two.

To Oleg, whose love and support made this possible.

Contents

List of Figures

List of Tables

Acknowledgments

I owe a special debt of gratitude to Professor Ellen Eisen. Her energy, curiosity, honesty, and unfailingly generosity with her time and expertise to those lucky enough to be her men tees, will always serve as an inspiration to me. I am also very appreciative to Mark van der Laan, Alan Hubbard, Sandrine Dudoit, Nicholas Jewell, Jack Colford, Jennifer Ahern, and Patrick Bradshaw, for their contributions to my intellectual growth. I feel truly privileged.

I would also like to thank the Centers for Occupational and Environmental Health (COEH) for the predoctoral support. I thank the epidemiological research group here in Environmental Health: Sally Picciotto, Dan Brown, Andreas Neophytou, and Sadie Costello for their generosity in sharing their time and expertise. Many, many thanks.

I thank all my colleagues and fellow students. I am especially thankful to Emily Yette, Alison Hughes, Jennifer Ames, Desya Levin, and Erika Garcia for their support and the many laughs we shared. I am thankful to my friends outside of school, Becky Hawrusik, Joel Cohen, Kyle Work, David Salazar and Dave Staconis, Eddie Herrador and Ryan Shriver, Jon Liew, Alex Betourné, Alex Loucks, David Hunter, and Tim Keller for being kind, supportive, and at times, a source of welcome distraction.

I am thankful to my mother Liza, and my father Bato. Throughout my life I have found strength in their tenacity, and endless comfort in their love.

Finally, I am grateful to my love, partner, and best friend, Oleg. I am grateful for his love, encouragement, patience and impatience, and the adventures we have shared. My best thoughts are of you.

Chapter 1

Introduction

This body of work is concerned with the estimation of causal effects of occupational exposures on incident cancer events. Many of the aspects of this work are motivated by a longitudinal study of approximately 40,000 workers in three automotive manufacturing facilities, the United Autoworkers - General Motors cohort study (UAW-GM). We demonstrate the ability of causal estimators to correct for time-varying confounding of the exposure-outcome relationship in an applied example that suggests a causal relationship between colon cancer and metalworking fluid exposure. We explore remaining barriers to unbiased estimation of causal parameters from cohorts of prevalent hires. We conclude with a discussion of classes of hypothetical of workplace interventions, and discuss factors that impact the behavior of causal parameters of such interventions.

Assessments of risks of workplace exposures are typically based on observational studies of occupational cohorts. Estimates from these studies are often subject to bias due to the healthy worker survivor effect (HWSE), a ubiquitous process that results in the healthiest workers accruing the most exposure over their lifetimes [1–5]. Workers in poorer health tend to accrue less exposure, whether by taking more time off, switching to lower exposed jobs, or by leaving the workforce entirely. The workers who remain active in the workforce and accrue the most exposure, are the healthiest ones. Time-varying confounding affected by prior exposure and selection bias have been identified as two potential sources of the HWSE. Time-varying confounding affected by prior exposure occurs if health status or other factors are on the causal pathway between earlier exposure and the outcome. While standard statistical models conditional on such factors result in biased estimates even under the null hypothesis, failure to adjust for them also results in bias [6].

Chapter 2 presents an analysis of the relationship between occupational exposure to straight, soluble, and synthetic metalworking fluids (MWFs) and incident colon cancer in the UAW-GM study. MWFs are a class of complex mixtures of chemicals, including several known or suspect carcinogens, that are aerosolized during marching operations and may pose a cancer risk to millions of manufacturing workers. Prior studies of the relationship between MWFs and colon cancer incidence have reported conflicting results [7, 8], likely due to the use of standard statistical methods that fail to adjust for time-varying confounding affected by prior exposure. We used longitudinal targeted minimum-loss based estimation (TMLE) to estimate the causal effect of exposure to each fluid type. Along with crude Kaplan-Meier survival estimates, we present adjusted survival curves predicting the estimated cumulative incidence of colon cancer among all workers had they been continually exposed above and below exposure cutoffs while actively employed. A detailed description of the steps involved in the analyses is also provided, intended to guide an epidemiological audience. This is the first applications of the TMLE framework in the field of cancer epidemiology.

Selection bias is a second component of the HWSE that occurs when the set of workers included into the study represents a sample of survivors drawn from a target population of all workers initially hired in a plant or industry. It operates through two main processes. The first occurs when follow-up ends with employment termination [9]. Bias may be induced by conditioning on employment status, a phenomenon known as collider stratification [10–12]. Selection bias may also be induced in the presence of differences in the susceptibility of workers to an occupational exposure: workers who are more susceptible to the adverse effects of an exposure experience the outcome earlier than the less susceptible [13]. Exposure effect estimates generally attenuate with follow-up time as workers who are more susceptible leave the cohort. Effect modification by susceptibility is responsible for left truncation bias in occupational cohorts of prevalent hires [14, 15]. It may also induce bias in the presence of *left filtering*, a less common feature of time-to-event data than censoring or truncation. Left filtering occurs when the reporting of secondary health outcomes begins after the original start of follow-up, and events that occur prior to the start of reporting are unknown. While cancer mortality follow-up started as early as 1941 in the UAW-GM study, follow-up for cancer incidence started in 1985, when the Michigan Cancer Registry was established. Studies of incident cancers in the UAW-GM cohort have been based on the sub-cohort of workers who were alive in 1985 [16–20]. Incident cancers are left filtered among this sub-cohort, since workers diagnosed before 1985 are included in the follow-up, but cannot become reported primary cancer cases after 1985.

Motivated by the left filtering of incident cancer events in the UAW-GM cohort, Chapter 3 presents a simulation study that assesses methods for the analysis of time-to-event data in the presence of left filtering, in the context of HWSE. We formally define left filtering as a process distinct from left truncation and left censoring. Five simulation scenarios are considered, each reflecting a key aspect of occupational cohorts. Complete (full) data where all incident cancers were known are simulated along with observed (left filtered) data. Incident cancers before the registry are unknown in the observed data. The exposure effect is measured as the difference in mean survival of workers always exposed versus never exposed to an occupational hazard while at work. Survival in the observed data is estimated using a novel delayed-entry adjusted Kaplan-Meier estimator that combines two known approaches:

the adjusted Kaplan-Meier estimator that adjusts for time-varying confounding [21], and a delayed-entry approach traditionally used to address random left truncation [22]. Bias for each of the five scenarios was measured as the difference between exposure effects in the full data, and the mean of exposure effects estimated in 500 observed datasets. We provide details on the estimation approaches and argue that the proposed approach may be appropriate in any study with a secondary outcome follow-up imposed on an existing cohort.

In Chapters 2 and 3 we consider causal effects of *dynamic* (realistic) interventions that assign exposure as a function of a worker's active employment status. To assess the etiologic effects of long-term exposure, causal effects can be further defined as contrasts of the distribution of counterfactual outcomes under *static* (etiologic) interventions that assign exposure and prevent leaving work throughout the study period. The two classes of interventions have different implications for disease prevention. In contrast to etiologic interventions that aim to estimate the biologic effect of long-term, sustained exposure, realistic interventions aim to estimate the expected disease experience of a population under a proposed standard. Occupational exposure guidelines are based on estimates of risks of long-term exposures, which correspond to causal effects of etiologic interventions that prevent leaving work. When the estimation of such effects is unfeasible, risk estimates of realistic interventions may be used to inform policy. It is therefore important to understand under what conditions the two effect measures are comparable, and identify factors that drive differences between the two. In Chapter 4 we present a simulation study that assesses the relationship between the two exposure-response measures in the context of the HWSE.

Chapter 5 concludes the dissertation by reviewing implications of the three studies conducted and priorities for future research.

Chapter 2

Metalworking Fluids and Colon Cancer Risk: Longitudinal Targeted Minimum-loss Based Estimation

2.1 Introduction

Despite the increased screening and improved treatment, colorectal cancer remains the third highest incident and fatal cancer worldwide [23, 24]. The rapid increase in the rates of colorectal cancer among migrants from low-risk to high-risk areas indicates that much of the disease burden is due to environmental causes. Many environmental carcinogens were initially identified in populations of highly exposed workers. Metalworking fluids (MWFs) are a class of complex mixtures used as coolants, lubricants, and anti-corrosives used during the fabrication of metal products in manufacturing industries that perform machining operations [25]. MWFs are aerosolized when sprayed, generating airborne particulate matter (PM) that has been linked to a number of cancers. With an estimated 4.4 million U.S. workers exposed in 1997 [26], and many more worldwide, MWF exposure poses a potential cancer hazard to workers in electronics manufacturing, new technologies, and alternative energy. The potentially carcinogenic nature of MWFs and their additives has long been noted [27]. MWFs have been linked to excess prostate cancer mortality [20], as well as excess incidence of laryngeal cancer [17, 28], bladder cancer [29], and malignant melanoma [16]. The few studies of MWFs in relation to incident colon cancer have reported conflicting findings. While an aerospace cohort study reported a statistically insignificant protective effect of mineral oils on colorectal cancer incidence [8], a Swedish population-based case-control study reported an elevated risk of incident colon cancer among male petrol station/automobile repair workers exposed to cutting fluids/oils after adjusting for known risk factors such as diet, physical activity and body mass [7]. A previous analysis of the United Autoworkers-General Motors (UAW-GM) cohort reported no association between straight, soluble, and synthetic

MWFs and incident colon cancer [18]. An important limitation of those studies is the use of standard statistical methods that fail to account for time-varying confounding affected by prior exposure. Time-varying confounding occurs if some factor is both a confounder of the exposure-outcome relationship and lies on the causal pathway between earlier exposure and the outcome. Active employment status, for instance, may both be affected by past exposure and predict future exposure. In addition, employment status may also be associated with the outcome of interest if a proportion of workers terminated employment because they were ill [5].

In the current work, we estimate the difference in the risk of colon cancer under hypothetical interventions to limit the average time weighted daily exposure in each year of employment, adjusting for time-varying confounding on the causal pathway. We used a novel approach, longitudinal targeted minimum-loss based estimation (TMLE), which has a number of advantages in the estimation of causal effects. To our knowledge, this is one of first applications of the longitudinal TMLE in the field of occupational epidemiology, and the first in the field of cancer epidemiology.

2.2 Methods

2.2.1 Study Population

The UAW-GM cohort study was initiated in 1984, jointly funded by the United Autoworkers labor union as well as the General Motor's management, to address workers' health concerns. The cohort includes 46,316 hourly workers from three automobile manufacturing plants in Michigan [27]. All hourly employees who had worked at least 3 years between January 1, 1938 and January 1, 1985 were included with the exception of employees who had ever worked in the large forge operation and may have been exposed to a variety of known carcinogenic agents [27]. Date of birth, sex, race, as well as the complete work history, including job title, department, and dates worked, were abstracted from employment records. Information on race was recorded for most worker, but remains unknown for approximately 10% of participants [30].

2.2.2 Outcome

Follow up for mortality began in 1941 and was recently extended to 2009. Vital status was obtained through the Social Security Administration, the National Death Index, as well as plant records and copies of state mortality files provided by the United Autoworkers Union [30]. Cohort members alive on January 1, 1985 ($N = 33,915$) constitute the UAW-GM cancer incidence sub-cohort. The analyses presented here are based on the cancer incidence sub-cohort, which was linked with the Michigan Cancer Registry to obtain cases diagnosed between January 1, 1985 and December 31, 2009. The diagnoses were classified using the International Classification of Diseases for Oncology, Third Edition (ICD-O-3). Notably, findings from previous analyses suggest that the effects of MWFs and their components may vary across the different regions of the large bowel [18, 31]. The analyses presented here are based on the first primary diagnosis of incident colon cancers (ICD-O-3 codes C18.0-C18.9), since the small number of rectal cancers (n=144) did not permit the evaluation of this sub-site.

2.2.3 Exposure Assessment

The three plants in this study represented exposure to three broad MWF classes: straight mineral oil, soluble oil (includes semisynthetic oils), and synthetic fluids. Quantitative exposure levels for each class of fluids had been previously developed for each time period, plant, department, and job-specific exposure category [32, 33]. Scale factors for each fluid, operation, and time period, were developed from a statistical model based on 394 air measurements collected by the company between 1958 and 1987 and determinants of exposure. The scale factors were then applied to plant, operation, and fluid-specific exposure levels from measurements collected by the original research study team in 1986-1987 [33]. An industrial hygienist revisited the plants in 1995 to update the scale factors for the 1985-1995 time period on the basis of company measurements collected since the study team's field visits [34].

Twenty-two percent of the UAW-GM cancer incidence sub-cohort was missing some of their work history. Subjects missing more than 50% of their work history were excluded (2.4%). Among the rest, missing work history information was interpolated by averaging the exposures from the previous and subsequent job for each subject. We used the annual quantitative measurements to compute an annual measure of exposure to each fluid type. Cumulative exposure was estimated by summing across years of employment. To account for a latency period, exposures were lagged by 15 years.

2.2.4 Statistical methods

We applied a targeted minimum-loss based estimation (TMLE) approach to estimate the cumulative incidence of colon cancer in each year t of follow-up in a cohort following a longitudinal exposure regimen specified by the researcher [35]. We considered regimens that set exposure above or below a cutoff while at work, and prevent right censoring by death. We used a dichotomous definition of exposure in which MWF levels above the cutoff were defined as "exposed", while MWF levels below the cutoff were defined as "unexposed". The cutoffs were determined a priori at the median exposure to each type of fluid among the colon cancer cases for the years during which they had non-zero exposure levels. We compared the estimated cumulative incidence of incident colon cancer if the worker population had been always exposed above the cutoff while actively employed, to the estimated cumulative incidence if the same worker population had been exposed below the cutoff (unexposed) while employed.

The description and the derivation of the statistical properties of the TMLE for multiple time point interventions are provided elsewhere [36], and a more detailed explanation of the assumed data structure, statistical models, identifiability assumptions, and target parameters are provided in next section. Briefly, the TMLE [37] is a semi-parametric substitution estimator that uses the efficient estimating equation framework [38–40], adjusts for time-varying confounding, is consistent under partial model misspecification (double robustness property), and is efficient when models for exposure/censoring mechanism and outcome are correctly specified (local efficiency). It involves estimation of two components: (1) the probability of being exposed (exposure mechanism) and remaining uncensored (censoring mechanism) conditional on covariates, and (2) the average outcome conditional on exposure and covariates (outcome model).

Exposure models were fit among actively employed workers. In addition to year of hire, sex, race, and year of follow-up, exposure models included a set of 15-year lagged time-varying covariates consisting of age and duration of employment at the start of t, as well as the proportion of the year spent on leave, an indicator of the plant at which the worker was employed, and cumulative exposure to each fluid type in the previous year (year $t-1$). Censoring models included year of hire, sex, race, age, year of follow-up, lagged duration of employment, the proportion of the year spent on leave, a plant indicator, cumulative exposure to each fluid type, and active employment status, all measured in year t. Both the exposure and censoring models were estimated using *SuperLearner*, a machine-learning approach that avoids the potential bias caused by assuming a fixed parametric relationship between covariates [41, 42]. SuperLearner uses cross-validation to create the best combination of algorithm specific estimates from a library of algorithms. In this application, the library we chose included Generalized linear models (GLMs) with Elastic Net Regularization [43], and Gradient Boosting Machines (GBM) with regression and classification trees [44]. We considered an ensemble of six elastic net models with parameter alpha ranging from 0 to 1, and optimal lambda picked by minimizing the cross validation mean-squared error. The tuning parameters for GBM were selected via the random grid search over the space of model parameters. For each year, model-based predicted exposure and censoring probabilities were used to estimate weights defined by inverse propensity scores of remaining uncensored and following the exposure of interest. Weights greater than 50 were set to 50. The truncation of the weights affected approximately 0.9%, 1.5%, and 0.4% of the observations in the analyses of the effects of straight, soluble, and synthetic MWFs, respectively.

Among the same population, we then applied the iterative g-computation formula [38], estimating an iterative series of conditional regressions predicting the average outcome at year t. Outcome models were estimated using main term logistic regression. While current exposure to each MWF was considered separately, models were adjusted for cumulative exposure to the other two by the end of the previous year. The initial estimator of the outcome regression was then updated in a targeting step that augments the initial outcome regression with information in the estimated exposure and censoring mechanisms. The targeting ensures that if either the treatment and/or the outcome models are correctly specified, the resulting estimator is efficient, in that it has the lowest variance among unbiased estimators in the model. The end result is a series of estimates of the cumulative incidence of colon cancer in each year t and exposure group. These estimates were used to calculate risk differences comparing exposures over the cutoff to under the cutoff. In addition, cumulative incidence estimates for each exposure group were used to create curves that indicate the estimated survival (1 - cumulative incidence) in each year. TMLE provides estimates of cumulative incidence at each time point considering all the past. Notably, since the estimated cumulative incidence at year $t+1$ is not constrained to be greater than cumulative incidence at year t, the TMLE-derived discrete survival curve is not necessarily a monotonically decreasing function. We present crude Kaplan-Meier survival curves in addition to TMLE curves for comparison.

The analysis was performed using the *stremr* [45] package in R [46].

2.2.5 The observed data structure

As previously noted, in the UAW-GM cancer incidence sub-cohort data were collected each year and were subject to right censoring. The observed data on each worker consists of measurements on baseline covariates, denoted by $L(0)$. In addition, for each worker the observed data includes annual measurements of 15-year lagged exposure to each metalworking fluid, the outcome, and confounding variables, starting in 1985, until each worker's end of follow-up at their death or 2009, whichever occurred first. The maximum observed follow-up was 25 years. The year when the worker's follow-up ends is denoted by $\tilde{T}$, and is defined as the earliest time to an incident colon cancer diagnosis denoted by T, or right censoring, denoted by C. At each year $t = 0, \ldots, \tilde{T}$, the worker's exposure above a predetermined cutoff to each type of a metalworking fluid is represented by the binary variable $E(t)$. $D(t)$ denotes a worker's right censoring indicator at year t. The combination $A(t) = (E(t), D(t))$ is referred to as the action at year t. At each year $t = 0, \ldots, \tilde{T}$ time-varying covariates such as intermittent time-off, years of employment, age, employment status, and cumulative exposure to each of the three fluids, are denoted by the multi-dimensional variable $L(t)$. $L(t)$ is defined from measurements that occur before the action at year t, $A(t)$, or are otherwise assumed to be unaffected by the actions at time t or thereafter. In addition to the aforementioned time-varying covariates, $L(t)$ includes an indicator of whether a colon cancer diagnosis occurred prior to the end of year t, $Y(t) = I(T < t) \in L(t)$. Furthermore, it is assumed that $Y(0)$ is constant 0 (the event of interest cannot occur at time $t = 0$). By definition, the outcome is missing at t if the worker was right censored at t. For notational simplicity, we use over-bars to denote covariate and exposure histories. For example, a subject's exposure history through time t is denoted by $\bar{E}(t) = (E(0), \ldots, E(t))$. We approach the observed data in this study as realizations of n independent and identically distributed (i.i.d.) copies of

$$O = \left(\bar{L}(t), \bar{A}(t) : t = 0, \ldots, \tilde{T}\right),$$

where the data on each unit i is denoted by O_i, for $i = 1, \ldots, n$.

2.2.6 Causal and statistical models

The probability distribution P_0 of O can be factorized according to the time ordering as

$$\begin{aligned} P_0(O) &= \prod_{k=0}^{K+1} P_0\left(L(k)|Pa(L(k))\right) \prod_{k=0}^{K} P_0\left(A(k)|Pa(A(k))\right) \\ &\equiv \prod_{k=0}^{K+1} Q_{0,L(k)}\left(O\right) \prod_{k=0}^{K} g_{0,A(k)}\left(O\right) \\ &\equiv Q_0\left(O\right) g_0\left(O\right), \end{aligned}$$

where $Pa(L(k) \equiv \left(\bar{L}(k-1), \bar{A}(k-1)\right)$ and $Pa(A(k) \equiv \left(\bar{L}(k), \bar{A}(k-1)\right)$ denote the parents of $L(k)$ and $A(k)$ in the time-ordered sequence, respectively. $Q_{0,L(k)}$ denotes the true conditional distribution of $L(k)$, given $Pa(L(k))$. $g_{0,A(k)} = g_{0,E(k)} g_{0,D(k)}$ denotes the true distribution of the treatment vector $(E(k), D(k))$ given $Pa(A(k))$. We define a statistical model $\mathcal{M}$ for the observed data distribution, P_0. If $\mathcal{Q}$ represents the set of all possible values for Q_0, and $\mathcal{G}$ represents the set of all possible values of g_0, then this statistical model can be represented as $\mathcal{M} = \{P = Q_g : Q \in \mathcal{Q}, g \in \mathcal{G}\}$. In this statistical model, $\mathcal{Q}$ puts no restrictions on the conditional distributions $Q_{0,L(k)}$, for $k = 0, \ldots, K+1$. Let

$$P^d(l) = \prod_{k=0}^{K+1} Q^d_{L(k)}\left(\bar{l}(k)\right),$$

where $Q^d_{L(k)}\left(\bar{l}(k)\right) = Q_{L(k)}\left(l(k)|\bar{l}(k-1), \bar{A}(k-1) = \bar{d}(k-1)\right)$, the so called G-computation formula for the post-intervention distribution correction with the intervention that sets each intervention node $A(t)$ to that determined by some rule $d(\bar{L}(t))$. In this study, we considered a class of dynamic regimes defined by a deterministic function $d(\bar{L}(t))$ of the observed data, where $d(\bar{L}(t))$ is used for setting the intervention nodes $A(t)$. In more detail, $d_{1,t}$ is a dynamic intervention that sets $E(t)$ to 1 at time t while a worker is actively employed, and sets $D(t)$ to 0. However, once the individual leaves work, $d_{1,t}(\bar{L}(t))$ then sets $E(t)$ and $D(t)$ to 0. Similarly, $d_{0,t}$ is defined as a dynamic intervention that sets $E(t)$ to 0 while a worker remains employed. Both interventions prevent right censoring. We define $\bar{d}_{\theta,t} = (d_{\theta,0}, \ldots, d_{\theta,t})$.

A causal model serves as the link between the observed data and the counterfactual data. We use the non-parametric structural equation model (NPSEM) framework [47, 48] to construct the following causal model, $\mathcal{M}^F$, for $k = 0, \ldots, K$:

$$\begin{aligned} L(k) &= f_{L(k)}\left(Pa(L(k)), U_{L(k)}\right) \\ A(k) &= f_{A(k)}\left(Pa(A(k)), U_{A(k)}\right), \\ &\ldots \\ L(K+1) &= f_{L(K+1)}\left(Pa(L(K+1)), U_{L(K+1)}\right), \end{aligned}$$

where $Pa(L(0))$ is null for convenience, $f_{A(k)}, f_{L(k)}$ are non-parametric deterministic functions, and $\left(U_{A(k)}, U_{L(k)}\right)$ are random unmeasured factors assumed to follow an unknown distribution, P_U. Under $\mathcal{M}^F$, each component of the data structure is generated as a deterministic function of its parents and the exogenous errors. Furthermore, this causal model can be used for generating the counterfactual values $L_d(k)$ under the dynamic treatment d_k, for $k = 0, \ldots, K$:

$$\begin{aligned}
A(k) &= f_{A(k)}\left(Pa(A(k)), U_{A(k)}\right) \\
L_d(k) &= f_{L(k)}\left(\bar{L}_d(k-1), d_{\theta,k}\left(\bar{L}_d(k-1)\right), U_{L(k)}\right) \\
&\quad ... \\
L_d(K+1) &= f_{L(K+1)}\left(\bar{L}_d(K), d_{\theta,K}\left(\bar{L}_d(K)\right), U_{L(K+1)}\right).
\end{aligned}$$

That is, we replace the intervention nodes in $f_{L(k)}$ with those set by our rule, and previous Y nodes by their previously generated counterfactual values. The counterfactual values of $Y_{\bar{d}_{\theta,t-1}}(t) \in L_{\bar{d}_{\theta,t-1}}(t)$ are then generated sequentially for each year. $Y_{\bar{d}_{\theta,t_0-1}}(t_0)$ for $\theta = 0, 1$ denotes a worker's potential outcome at time t_0 had she been exposed between study entry and time $t_0 - 1$ according to rule $\bar{d}_{\theta,t_0-1}$. Specifically, following the Neyman-Rubin model [49], for $\bar{d}_t \in (\bar{d}_{0,t}, \bar{d}_{1,t})$, the counterfactual $Y_{\bar{d}_{t-1}}(t)$ is the outcome an individual would have at time t if, possibly contrary to fact, they had exposure assigned according to rule $\bar{d}_{t-1}$. For notational convenience we will denote $Y_{\bar{d}_{t-1}}(t)$ as $Y_{\bar{d}}(t)$.

2.2.7 Identifiability

Interventions are defined with respect to counterfactual outcomes of interest. The probability distribution of the counterfactual $Y_{\bar{d}}(t)$ is called the post-intervention distribution of Y. Under the sequential randomization assumption and the positivity assumption, $Y_{\bar{d}}(t)$ and $Y^{\bar{d}}(t)$ have the same probability distribution. Formally, the assumptions required for identifiability are as follows:

Sequential randomization assumption.

$$A(k) \perp\!\!\!\perp Pa(A(k)) \text{ for } k = 1, \ldots, K.$$

Positivity assumption.

$$P_0\left(A(k) = d_k(\bar{L}(k)) | \bar{L}(k), \bar{A}(k-1) = \bar{d}_k(\bar{L}(k-1))\right) > 0 \text{ almost everywhere.}$$

Informally, the sequential randomization assumption states that at each time point k, all common causes of the L and A nodes are measured and included in the dataset. The positivity assumption states that the probability that all workers follow an intervention determined according to rule d_k for $k = 1, \ldots, K$ is positive. In other words, the positivity assumption states that some workers will experience the outcome of interest for all strata of the covariates in the observed data.

2.2.8 Target parameters

As first demonstrated by Robins [50], under the sequential randomization and positivity assumptions, the intervention specific means $E_0(Y_{\bar{d}}(t))$ can be identified through a sequence of recursively defined conditional expectations, the first of which takes the form:

$$\bar{Q}^d_{L(t)} = E_0\left(Y(t)|\bar{L}(t-1), \bar{A}(t-1) = \bar{d}_{t-1}\left(\bar{L}(t-1)\right)\right).$$

This regression corresponds to the regression of $Y(t)$ on the past covariates and intervention nodes, performed among the population of treatment regimen followers, i.e., evaluated at the values of the intervention nodes that would have been assigned by applying the dynamic rule $\bar{d}_t$. The quantity $\bar{Q}^d_{L(t)}$ is then regressed in reverse chronological order on covariates and intervention nodes set by $\bar{d}_t$ up to time $t-2, t-3, \ldots, 0$. Specifically, for $t = k-1, \ldots, 1$:

$$\bar{Q}^d_{L(t)} = E_0\left(\bar{Q}^d_{L(t+1)}|\bar{L}(t-1), \bar{A}(t-1) = \bar{d}_{t-1}\left(\bar{L}(t-1)\right)\right).$$

When $k = 0$ the results is a final constant $E_0(Y_{\bar{d}}(t)) = \bar{Q}^d_{L(0)} = E_0\left(\bar{Q}^d_{L(1)}|L(0)\right)$. Under the stated assumptions, the distribution of the counterfactual outcome $Y_{\bar{d}}(t)$ is equal to the distribution of the observed outcome under intervention, $\bar{Q}^{\bar{d}}_{L(0)}$, which is estimated from the observed data [36, 51].

The causal parameter of interest is the cumulative incidence of the outcome for each regimen $\bar{d}_t$ is given by $P\left(Y_{\bar{d}}(t_0+1) = 1\right)$. The parameter of interest is then defined as the difference between the cumulative incidences at time t_0 associated with the regimens $\bar{d}_{1,t_0}$ and $\bar{d}_{0,t_0}$:

$$\psi^{RD} = P\left(Y_{\bar{d}_1}(t_0+1) = 1\right) - P\left(Y_{\bar{d}_0}(t_0+1) = 1\right).$$

2.3 Results

Our analytic cohort consisted of 33,063 workers. During the 25-year follow-up we identified 466 incident colon cancers. *Table 2.1* compares the characteristics of the study population at baseline, in 1985. Colon cancer cases were more likely to be women, black, and older at the time of their hire than non-cases. Approximately 8% of the cases were actively employed at the time of their colon cancer diagnosis. The mean age at diagnosis was 68.2 years (SD = 10.3 years). Workers diagnosed with colon cancer had a shorter mean follow-up, longer mean duration of employment, and were more likely to work in Plant 1 at baseline compared to their healthy counterparts. In addition, workers diagnosed with colon cancer were more likely to have been exposed at baseline, and had higher cumulative exposure to all three fluid types compared to non-cases.

	Incident Colon Cancer Cases	Non-cases
N	466	32597
Person-year contribution, 1985 – 2009	6185	685850
Censored (Death), n (%)	–	35.25
Duration of follow-up, *mean (SD)*	13.27 (6.91)	21.04 (6.81)
Age at hire, *mean (SD)*	32.75 (9.13)	29.25 (7.98)
Female, (%)	14.16	13.58
Black, (%)	25.11	18.77
Age at diagnosis, *mean (SD)*	68.21 (10.3)	–
Actively employed at diagnosis, (%)	7.73	–
Covariates at baseline (1985)		
Age, *mean (SD)*	55.94 (11.63)	46.24 (14.07)
Duration of employment, *mean (SD)*	18.19 (10.08)	14.03 (8.69)
Active at work, (%)	46.78	58.71
Active workers in plant 1, (%)	29.82	18.43
Active workers in plant 2, (%)	30.28	40.67
Active workers in plant 3, (%)	39.91	40.84
Ever exposed, (%)		
Straight	41.85	28.46
Soluble	70.17	48.37
Synthetic	18.24	13.51
Cumulative exposure among exposed person-years ($\frac{mg}{m^3} year$), *median (IQR)*		
Straight	1.33 (0.36 - 6.35)	0.93 (0.26 - 3.79)
Soluble	8.63 (2.77 - 18.3)	5.90 (2.13 - 14.74)
Synthetic	1.43 (0.29 - 3.54)	0.89 (0.24 - 2.42)
Annual Exposure ($\frac{mg}{m^3}$) during exposed person years of follow-up, *median (IQR)*		
Straight	0.10 (0.04 - 0.36)	0.07 (0.03 - 0.32)
Soluble	0.37 (0.18 - 0.60)	0.27 (0.16 - 0.47)
Synthetic	0.04 (0.02 - 0.13)	0.04 (0.02 - 0.12)

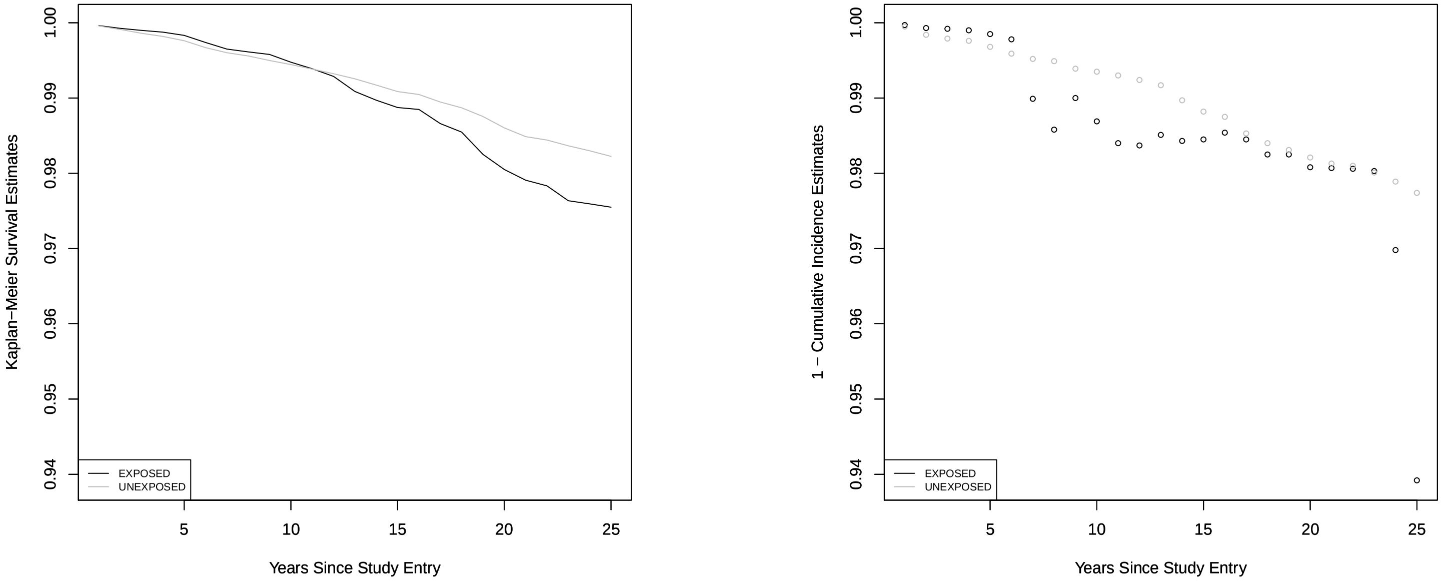
Table 2.1: Characteristics of the United Autoworkers-General Motors (UAW-GM) study population.

Crude Kaplan-Meier and TMLE survival curves comparing the experience of workers under the two exposure regimens for each type of MWF are presented in *Figures 2.1a-2.3b*. Kaplan-Meier survival estimates for all three fluid types (*Figures 2.1a, 2.2a, 2.3a*) indicate similar survival among exposed and unexposed workers during the first few years of follow-up, with survival among workers exposed to synthetic MWFs exceeding that of unexposed workers. Exposed workers had poorer survival for all three fluid types during the second half of the follow-up, and the difference in survival increased over time. Of note, while both crude and TMLE estimates indicate poorer survival for exposed workers, adjustment for time-varying confounding via TMLE resulted in greater survival differences (*Figures 2.1b, 2.2b, 2.3*b).

Table 2.3 presents, the number of workers at risk at the start of the year each year, as well as the number of workers that were right censored (died) and diagnosed with colon cancer during that year. The risk differences comparing the cumulative incidence of colon cancer as predicted for a cohort exposed above the cutoff to the cumulative incidence for that same cohort exposed below the cutoff while at work are also presented in *Table 2.3.* Cumulative incidences for each exposure group and year of follow-up are provided in *Table 2.2.* Overall, exposure to **straight** MWFs resulted in higher cumulative incidence of colon cancer as reflected by positive risk difference estimates. Estimates reached statistical significance in the last year of follow-up. Specifically, we estimated that the 25-year cumulative incidence of colon cancer would be approximately 3.80% higher if workers were always exposed above the $0.100\frac{mg}{m^3}$ cutoff than if the same workers had always been exposed below the cutoff while at work (RD = 0.038, 95% CI = 0.022 to 0.054).

Exposure to **soluble** MWFs was also associated with increased risk of colon cancer. We estimated that the 25-year risk difference comparing a cohort of workers always exposed above the $0.369\frac{mg}{m^3}$ cutoff, to the same cohort of workers always exposed below the cutoff while at work was 0.002 (95% CI = -0.016 to 0.012) (*Table 2.3*).

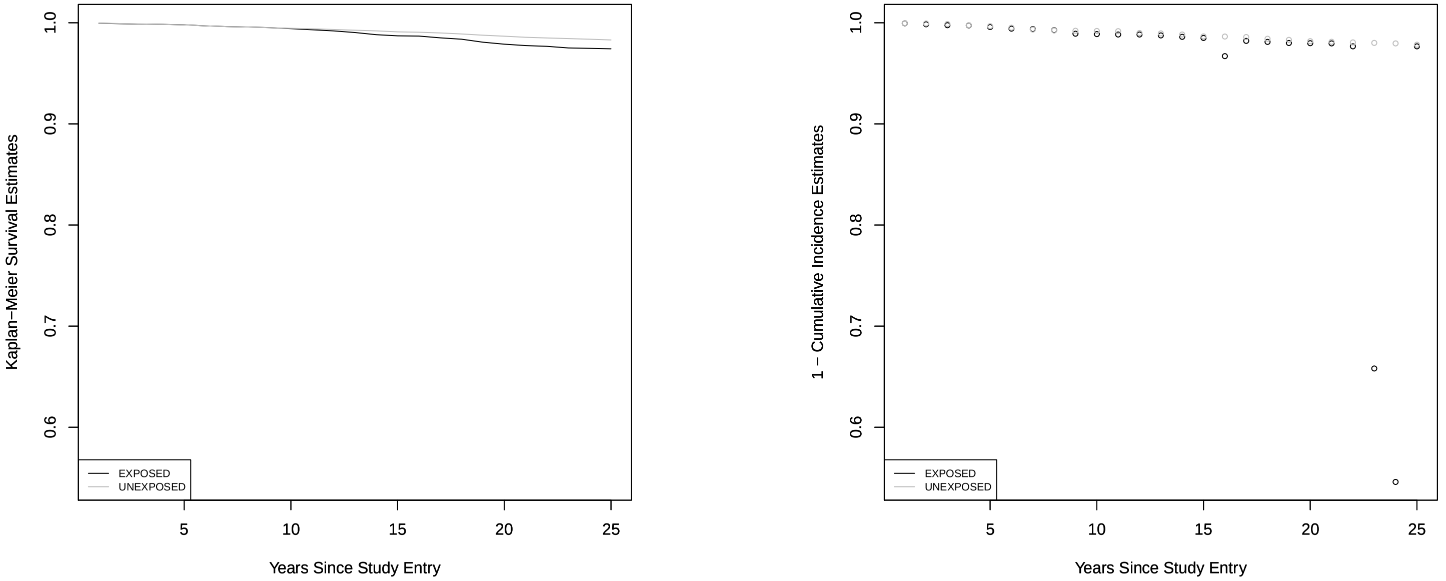
Exposure to **synthetic** MWFs was also associated with higher cumulative incidence of colon cancer among exposed workers compared to unexposed workers (*Table 2.3*). The estimated 25-year risk difference was 0.008 (95% CI = 0.002 to 0.014).



(a) Crude Kaplan-Meier Survival Estimates

(b) TMLE Estimates: 1 - Cumulative Incidence

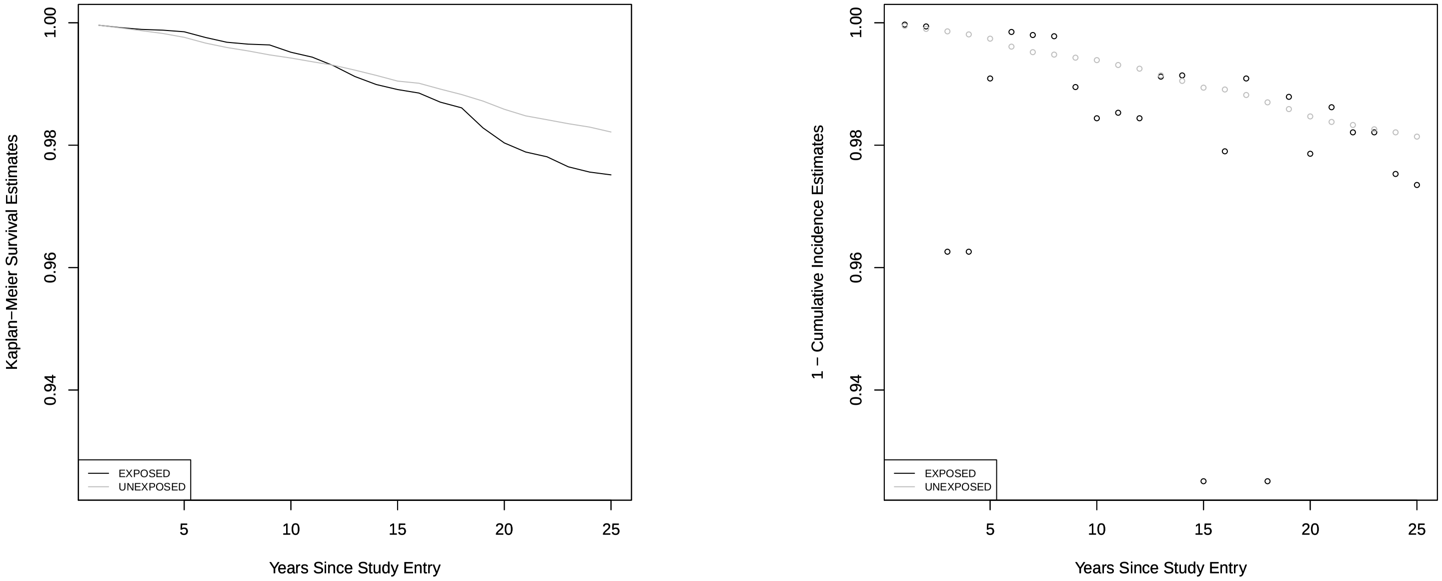
Figure 2.1: *Straight Metalworking Fluids.* Crude and TMLE estimated colon cancer survival, among cohorts of workers always exposed above and below the median while at work.



(a) Crude Kaplan-Meier Survival Estimates

(b) TMLE Estimates: 1 - Cumulative Incidence

Figure 2.2: *Soluble Metalworking Fluids.* Crude and TMLE estimated colon cancer survival, among cohorts of workers always exposed above and below the median while at work.



(a) Crude Kaplan-Meier Survival Estimates

(b) TMLE Estimates: 1 - Cumulative Incidence

Figure 2.3: *Synthetic Metalworking Fluids.* Crude and TMLE estimated Colon Cancer survival, among cohorts of workers always exposed above and below the median while at work.

2.4 Discussion

These results provide evidence that increased risk of incident colon cancer is associated with occupational exposure to straight, soluble, and synthetic MWFs after accounting for possible time-varying confounders on the causal pathway.

An aerospace cohort that only examined colon and rectal cancer incidence combined in relation to mineral oil exposure reported a statistically insignificant protective effect of mineral oils; however, water-based MWF were not examined [8]. The reasons for the inconsistencies between our findings and those in aerospace cohort are not clear, but suggest that, in addition to differences in the definition of the outcome, there may be important differences in the formulation of MWF types, processes, or work practices. While earlier mortality studies of the UAW-GM cohort found no associations with colon cancer mortality and MWF exposure [34], a previous analysis of the UAW-GM cancer incidence sub-cohort reported that soluble and synthetic MWF exposure were associated with a modest increase in the risk of colon cancer [18]. However when known carcinogenic components of soluble and synthetic MWFs were individually examined, both biocides (HR = 1.04, 95% CI: 1.02 - 1.07) and nitrosamines (HR = 1.02, 95% CI: 1.00 - 1.04) were found to increase the risk of colon cancer [18]. The presence of these components could describe the observed association of synthetic and soluble MWFs and colon cancer in our analyses.

A number of potential limitations should be considered in the interpretation of these findings. The time-varying job exposure matrix is the hallmark strength of the UAW-GM cohort study. However, the possibility for exposure misclassification exists. Few individual-level covariates were available for this cohort, raising concerns about residual unmeasured confounding. Dietary factors, physical inactivity, and excess body weight are important risk factors for colon cancer and would be confounders in this study if they were also associated with exposure. There is no a priori reason to expect that any of these factors are likely to vary with exposure. The assumption of causal consistency, which is subsumed by our approach [50], may be jeopardized by the categorization of exposure; it is unlikely that all workers would have the same outcome had they been assigned to any exposure level above the cutoff that defined the category. Data on the use of protective equipment is not available for this group. A last limitation of the proposed work is the assumption of a unique latency period of 15 year for all workers, when in fact latency periods may vary across individuals [52].

In addition to the aforementioned limitations this work has several strengths. A feature of the UAW-GM cohort is that follow-up continues past employment termination. As unemployed workers cannot be exposed, the data contains subject-times in which the probability of exposure is 0, resulting in a violation of the positivity assumption. This has been used as a justification for the use of G-estimation of accelerated failure time models in occupational epidemiology [53, 54], which do not rely on the positivity assumption. However,

Year of Follow-up	Straight MWFs		Soluble MWFs		Synthetic MWFs	
	$\geq 0.100\frac{mg}{m^3}$	$< 0.100\frac{mg}{m^3}$	$\geq 0.369\frac{mg}{m^3}$	$< 0.369\frac{mg}{m^3}$	$\geq 0.039\frac{mg}{m^3}$	$< 0.039\frac{mg}{m^3}$
1	0.000 (0.000 - 0.001)	0.001 (0.000 - 0.001)	0.001 (0.000 - 0.001)	0.000 (0.000 - 0.001)	0.000 (0.000 - 0.001)	0.001 (0.000 - 0.001)
2	0.001 (0.000 - 0.001)	0.002 (0.000 - 0.004)	0.002 (-0.001 - 0.004)	0.001 (0.000 - 0.001)	0.001 (0.000 - 0.001)	0.001 (0.001 - 0.001)
3	0.001 (0.001 - 0.001)	0.002 (0.000 - 0.004)	0.002 (-0.002 - 0.007)	0.001 (0.001 - 0.002)	0.037 (0.031 - 0.044)	0.001 (0.001 - 0.002)
4	0.001 (0.001 - 0.001)	0.002 (0.000 - 0.005)	0.003 (-0.002 - 0.007)	0.002 (-0.002 - 0.007)	0.037 (0.031 - 0.044)	0.002 (0.001 - 0.002)
5	0.002 (0.001 - 0.002)	0.003 (0.000 - 0.007)	0.004 (-0.002 - 0.010)	0.003 (-0.002 - 0.009)	0.009 (0.005 - 0.013)	0.003 (0.002 - 0.003)
6	0.002 (0.002 - 0.003)	0.004 (0.001 - 0.008)	0.006 (-0.001 - 0.013)	0.005 (-0.001 - 0.011)	0.002 (0.001 - 0.002)	0.004 (0.003 - 0.005)
7	0.010 (0.005 - 0.015)	0.005 (0.001 - 0.008)	0.006 (-0.001 - 0.013)	0.007 (-0.001 - 0.014)	0.002 (0.002 - 0.003)	0.005 (0.004 - 0.006)
8	0.014 (0.009 - 0.020)	0.005 (0.002 - 0.009)	0.007 (-0.001 - 0.015)	0.008 (-0.001 - 0.016)	0.002 (0.002 - 0.003)	0.005 (0.004 - 0.006)
9	0.010 (0.005 - 0.016)	0.006 (0.001 - 0.011)	0.011 (0.001 - 0.021)	0.008 (0.000 - 0.016)	0.011 (0.006 - 0.015)	0.006 (0.004 - 0.007)
10	0.013 (0.007 - 0.019)	0.007 (0.002 - 0.011)	0.011 (0.001 - 0.021)	0.008 (0.000 - 0.016)	0.016 (0.010 - 0.021)	0.006 (0.005 - 0.007)
11	0.016 (0.010 - 0.022)	0.007 (0.002 - 0.012)	0.012 (0.002 - 0.022)	0.008 (0.000 - 0.016)	0.015 (0.009 - 0.020)	0.007 (0.006 - 0.008)
12	0.016 (0.010 - 0.023)	0.008 (0.003 - 0.012)	0.012 (0.002 - 0.022)	0.010 (0.001 - 0.019)	0.016 (0.010 - 0.021)	0.008 (0.006 - 0.009)
13	0.015 (0.009 - 0.021)	0.008 (0.004 - 0.013)	0.013 (0.002 - 0.023)	0.011 (0.001 - 0.020)	0.009 (0.005 - 0.013)	0.009 (0.007 - 0.010)
14	0.016 (0.010 - 0.022)	0.010 (0.004 - 0.017)	0.014 (0.003 - 0.024)	0.012 (0.002 - 0.021)	0.009 (0.005 - 0.012)	0.010 (0.008 - 0.011)
15	0.016 (0.010 - 0.021)	0.012 (0.005 - 0.019)	0.015 (0.004 - 0.026)	0.014 (0.003 - 0.024)	0.075 (0.068 - 0.082)	0.011 (0.009 - 0.013)
16	0.015 (0.008 - 0.021)	0.013 (0.005 - 0.020)	0.033 (0.018 - 0.048)	0.014 (0.003 - 0.024)	0.021 (0.015 - 0.027)	0.011 (0.009 - 0.013)
17	0.016 (0.009 - 0.022)	0.015 (0.006 - 0.023)	0.018 (0.007 - 0.029)	0.014 (0.004 - 0.025)	0.009 (0.005 - 0.013)	0.012 (0.010 - 0.014)
18	0.018 (0.011 - 0.024)	0.016 (0.007 - 0.025)	0.019 (0.007 - 0.031)	0.016 (0.005 - 0.027)	0.075 (0.067 - 0.082)	0.013 (0.011 - 0.015)
19	0.018 (0.010 - 0.025)	0.017 (0.008 - 0.026)	0.020 (0.008 - 0.032)	0.017 (0.006 - 0.028)	0.012 (0.008 - 0.016)	0.014 (0.012 - 0.016)
20	0.019 (0.012 - 0.026)	0.018 (0.009 - 0.027)	0.020 (0.008 - 0.032)	0.018 (0.007 - 0.030)	0.021 (0.003 - 0.040)	0.015 (0.013 - 0.018)
21	0.019 (0.012 - 0.027)	0.019 (0.010 - 0.027)	0.020 (0.008 - 0.033)	0.019 (0.007 - 0.030)	0.014 (0.010 - 0.018)	0.016 (0.014 - 0.019)
22	0.019 (0.012 - 0.027)	0.019 (0.010 - 0.028)	0.023 (0.011 - 0.036)	0.019 (0.008 - 0.031)	0.018 (0.012 - 0.024)	0.017 (0.014 - 0.019)
23	0.020 (0.013 - 0.027)	0.020 (0.011 - 0.029)	0.342 (0.324 - 0.360)	0.020 (0.008 - 0.032)	0.018 (0.012 - 0.024)	0.017 (0.015 - 0.020)
24	0.030 (0.022 - 0.038)	0.021 (0.012 - 0.030)	0.454 (0.433 - 0.475)	0.020 (0.009 - 0.032)	0.025 (0.020 - 0.030)	0.018 (0.015 - 0.020)
25	0.061 (0.048 - 0.074)	0.023 (0.013 - 0.032)	0.023 (0.011 - 0.036)	0.022 (0.010 - 0.034)	0.027 (0.021 - 0.032)	0.019 (0.016 - 0.021)

Table 2.2: Estimated cumulative incidence (95% CI) of colon cancer for workers exposed at the indicated level of each metalworking fluid.

Year of Follow-up	N at Risk	N Censored	N Cases	Estimated Risk Difference (95% CI)		
				Straight MWFs	Soluble MWFs	Synthetic MWFs
1	33063	314	13	0.000 (0.000 - 0.000)	0.000 (0.000 - 0.000)	0.000 (0.000 - 0.000)
2	32736	374	15	-0.001 (-0.003 - 0.001)	0.001 (-0.002 - 0.004)	0.000 (-0.001 - 0.000)
3	32347	384	15	-0.001 (-0.003 - 0.001)	0.001 (-0.003 - 0.005)	0.036 (0.030 - 0.042)
4	31948	423	14	-0.001 (-0.004 - 0.001)	0.000 (-0.006 - 0.006)	0.036 (0.029 - 0.042)
5	31511	391	19	-0.002 (-0.005 - 0.002)	0.001 (-0.007 - 0.009)	0.007 (0.002 - 0.011)
6	31101	410	28	-0.002 (-0.006 - 0.002)	0.001 (-0.008 - 0.010)	-0.002 (-0.003 - -0.002)
7	30663	400	23	0.005 (-0.001 - 0.011)	-0.001 (-0.011 - 0.010)	-0.003 (-0.004 - -0.002)
8	30240	439	17	0.009 (0.003 - 0.016)	0.000 (-0.012 - 0.011)	-0.003 (-0.004 - -0.002)
9	29784	477	19	0.004 (-0.003 - 0.011)	0.003 (-0.010 - 0.016)	0.005 (0.000 - 0.010)
10	29288	457	14	0.007 (-0.001 - 0.014)	0.003 (-0.010 - 0.016)	0.010 (0.004 - 0.015)
11	28817	462	16	0.009 (0.001 - 0.017)	0.003 (-0.010 - 0.016)	0.008 (0.002 - 0.014)
12	28339	458	17	0.009 (0.001 - 0.017)	0.002 (-0.012 - 0.015)	0.008 (0.002 - 0.014)
13	27864	485	20	0.007 (-0.001 - 0.014)	0.002 (-0.012 - 0.016)	0.000 (-0.004 - 0.004)
14	27359	464	22	0.005 (-0.003 - 0.014)	0.002 (-0.012 - 0.017)	-0.001 (-0.005 - 0.003)
15	26873	481	23	0.004 (-0.005 - 0.013)	0.002 (-0.014 - 0.017)	0.064 (0.057 - 0.072)
16	26369	467	9	0.002 (-0.008 - 0.012)	0.019 (0.001 - 0.038)	0.010 (0.004 - 0.017)
17	25893	523	23	0.001 (-0.010 - 0.012)	0.004 (-0.012 - 0.019)	-0.003 (-0.007 - 0.001)
18	25347	553	20	0.002 (-0.009 - 0.012)	0.003 (-0.013 - 0.019)	0.062 (0.054 - 0.070)
19	24774	498	26	0.001 (-0.011 - 0.012)	0.003 (-0.013 - 0.020)	-0.002 (-0.006 - 0.002)
20	24250	437	32	0.001 (-0.010 - 0.013)	0.002 (-0.015 - 0.019)	0.006 (-0.012 - 0.025)
21	23781	545	23	0.001 (-0.011 - 0.012)	0.002 (-0.015 - 0.018)	-0.002 (-0.007 - 0.002)
22	23213	497	14	0.000 (-0.011 - 0.012)	0.004 (-0.013 - 0.021)	0.001 (-0.005 - 0.007)
23	22702	537	15	0.000 (-0.011 - 0.011)	0.322 (0.301 - 0.343)	0.001 (-0.006 - 0.007)
24	22150	514	13	0.009 (-0.003 - 0.021)	0.434 (0.410 - 0.458)	0.007 (0.001 - 0.012)
25	21623	502	16	0.038 (0.022 - 0.054)	0.002 (-0.016 - 0.019)	0.008 (0.002 - 0.014)

Table 2.3: Number of workers at risk, number censored, number of incident colon cancers, and the estimated the risk difference (95% CI) comparing workers exposed above and below the cutoff of fluid type while at work.

G-estimation is more appropriate for commonly occurring outcomes (e.g., heart disease or all-cause mortality) compared to relatively rare events such as incident cancers [54]. Instead, we used an approach that allows us to estimate causal parameters of dynamic interventions [55], interventions that assign exposure in response to a subject's covariate values, such as her employment status. The causal effects of such interventions have been assessed in few simulated or applied examples in the fields of cancer and occupational epidemiology [56–58]. The prospective design of the UAW-GM study precludes recall bias. Its large size and long period of follow-up provide ample statistical power. The availability of comprehensive quantitative exposure levels is one of the strengths of this analysis. Another strength is the availability of data on intermittent time-off work, which was used as a time-varying health surrogate. The use of cross-validated ensemble learners [41] improved model fit possibly reducing bias in comparison with main term logistic models. The small number of time-varying covariates resulted in limited positivity violations since few combinations of covariates were predictive of exposure status.

Our analysis is the first to support a possible causal relationship between MWFs, particularly straight fluids, and incident colon cancer. Given the ubiquity of exposure to these chemicals, lowering occupational limits may prevent a large number of colon cancers worldwide. By estimating the risk reduction associated with lowering occupational exposure limits for specific types of MWFs, we provide a public health framework for our findings.

Chapter 3

Drivers of Biased Effect Estimates in Left Filtered Data

3.1 Introduction

Longitudinal studies of occupational disease often follow participants prospectively, identifying events of interest as they occur. Ideally, all study participants would be enrolled at hire. However, in many occupational cohort studies the follow-up period begins long after workers were hired, subjecting these studies to bias from a number of sources. Left truncation occurs when individuals who have already experienced the event of interest are excluded from the study [59]. Left censoring occurs when a subject has experienced the event of interest prior to study entry, but its exact timing may be unknown [59]. Left filtering refers to a situation in which not only the timing, but even whether the event has occurred at all, is unknown. The latter is the case for incident cancers in the United Autoworkers – General Motors (UAW-GM) cohort study.

The details of the UAW-GM study have previously been described [16, 27]. Briefly, the cohort consists of approximately 46,000 workers hired from 1938 until 1982 in three automotive manufacturing plants in Michigan where workers were exposed to metalworking fluids (MWFs). MWFs are complex mixtures of mineral oils and chemicals, including several known or suspected carcinogens [16–18, 27–29]. Exposure and work history were recorded from hire until 1994. However, the reporting of incident cancers started in 1985, when the Michigan Cancer Registry was established, and ended in 2009. Studies of incident cancers in the UAW-GM cohort have included the sub-cohort of workers who were alive in 1985 [16–20]. Incident cancer events are left filtered among this sub-cohort, since workers who were diagnosed before 1985 are included in the follow-up, but not eligible to become primary incident cancer cases. In addition, the sub-cohort may represent a biased sample of the full cohort, particularly in the presence of the healthy-worker survivor effect (HWSE).

Time-varying confounding affected by prior exposure and selection bias have long been identified as two potential sources of the HWSE [60]. Time-varying confounding affected by prior exposure occurs if a covariate is both a confounder of the exposure-outcome relationship and lies on the causal pathway between earlier exposure and the outcome. Health status, for instance, may both be affected by past exposure and predict future exposure. It may also be associated with the outcome of interest if some workers terminated employment because they were ill [5]. Selection bias occurs because the sub-cohort alive in 1985 represents a sample of survivors drawn from a target population of all workers initially hired. One of the ways in which selection bias may perpetuate itself is if workers who are more susceptible to the adverse health effects of an exposure experience the outcome of interest, die, or self-select out of the workforce more rapidly than the less susceptible [13]. Effect estimates generally attenuate with follow-up time as the more susceptible workers leave the cohort [13]. Under this causal structure, it is conceivable that prior UAW-GM analyses based on the sub-cohort of workers alive in 1985 may have underestimated the true effect of MWF exposure. In addition, many of the previous analyses have used traditional methods, such as Cox regression, which rely on the conditional independence of the time to the event of interest and time to right censoring. Right censoring occurs when a participant has not yet experienced the event of interest at study end [59]. The latter assumption would be violated if workers with unreported (latent) cancer diagnoses were at higher risk of death.

A quantitative assessment of the bias introduced by analytical approaches in the presence of left filtering has not been carried out in the context of the healthy-worker survivor effect. In fact, an informal review of the literature yielded no previous studies proposing estimation approaches in analogous settings. Given that the National Institutes of Health (NIH) list over 50 disease registries in the U.S. alone, one might anticipate many studies with a secondary follow-up imposed on an existing cohort. To address this gap in the literature, we use a simulation study that assesses the impact of left filtering in workplace studies.

3.2 Methods

3.2.1 Data description and notation.

To understand how the exposure-response relationship estimated in the subset of workers alive in 1985 may differ from that in the full UAW-GM cohort, we simulated data in which incident events were observed from hire until death or end of follow-up. We refer to this underlying data structure as the full data or the full cohort, and denote it by X. A schematic of the full data is presented in *Figure 3.1A.* We additionally simulated the observed data O, in which (as in UAW-GM) incident cancers were reported only after the establishment of the cancer registry. In *Figure 3.1B* we present the underlying structure of the observed data. In *Figure 3.1C* we present the data we actually observe. We denote the time from hire until a cancer diagnosis, the event of interest, by T. C denotes the time to right censoring, defined as the time from hire till death or end of follow-up. The time from hire to the start of the cancer registry is denoted by C^*.

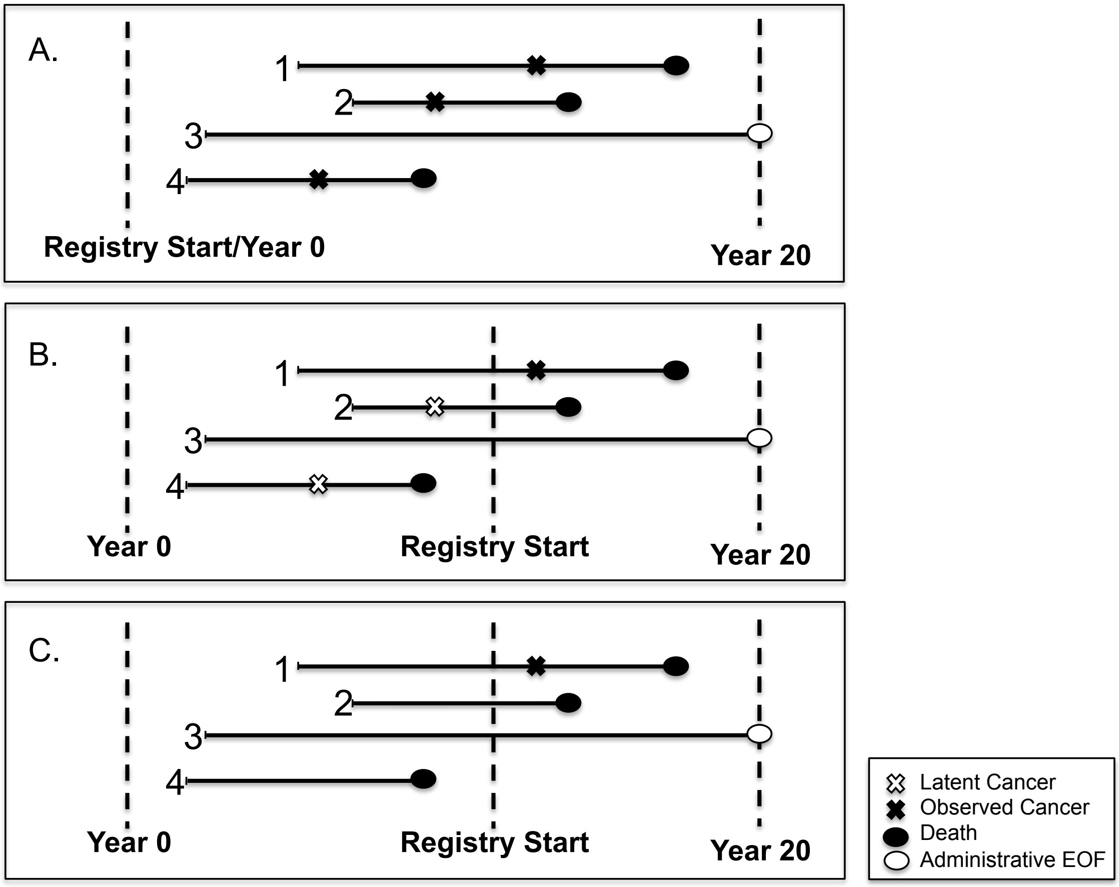


Figure 3.1: *Full data (A).* The registry started before all workers were hired, and all incident cancers (workers 1, 2, and 4) are known. Underlying structure of the *observed data (B)*: the registry started after all workers were hired, and latent cancers that occurred prior to its start (workers 2, and 4) are unknown. *The observed data (C)*: worker 1 is a known incident cancer case; workers 2 and 4 are considered non-cases.

For each worker we simulated a set of measured baseline covariates W, an indicator of susceptibility to exposure-related effects S, and a set of time-varying covariates. For each year t, we simulated $N(t)$: an indicator of active employment status; $H(t)$: an indicator of a diagnosis of an adverse health event, such as a chronic health condition; $E(t)$: the exposure node; and $D(t)$: an indicator of death (right censoring). Workers who terminated employment could not become actively employed at a later time. In addition, once diagnosed with an adverse health event, workers were assumed to experience it until the end of follow-up. $E(t)$ was set to 1 if a worker was exposed to the occupational hazard under study during year t, and 0 otherwise. $D(t) = I(t > C)$ was set to 1 if a worker died before the end of year t, and 0 otherwise. Next, we define $Y^*(t) = I(t \geq T)$ as the indicator of a cancer diagnosis (whether or not it was observed) on or before year t, and $Y(t)$ as an indicator of a cancer diagnosis reported in the registry (observed cancer diagnosis). Note that while $Y(t)$ and $Y^*(t)$ are always equivalent in the full data, in the observed data they are equivalent only if the cancer was first diagnosed after the start of the cancer registry (i.e., if $T \geq C^*$). A bar over a variable denotes the history from baseline up to year t. In sum, we assumed the following full data structure on n independent and identically distributed (iid) units in X:

$$X_i(t) = \left(C_i^*, W_i, S_i, \bar{H}_i(t), \bar{N}_i(t), \bar{E}_i(t), \bar{D}(t), \bar{Y}_i^*(t) = \bar{Y}_i(t)\right).$$

Since incident cancers were reported from hire, $C^* = 0$ for all subjects in the full data. The data structure on each unit in O is given by:

$$O_i(t) = (C_i^*, W_i, \bar{H}_i(t), \bar{N}_i(t), \bar{E}_i(t), \bar{D}_i(t), \bar{Y}_i(t)).$$

While $Y^*(t)$ is truly of interest, in the observed data we only get to measure its proxy, $Y(t)$. Of note, for some individuals i in the observed data we observe the exact failure time ($C_i^* < T_i \leq C_i$), while for others the failure time is right-censored ($C_i^* < C_i < T_i$). Since incident cancer events occurring prior the establishment of the registry are left filtered and the ones occurring after the end of follow-up are right censored, the observed data is both left filtered and right censored (i.e., $C_i^* < T_i < C_i$) [22].

3.2.2 Causal model.

We used the non-parametric structural equation model (NPSEM) framework [61, 62] to construct the following causal model, denoted as $\mathcal{M}^F$:

$$\begin{aligned}
C^* &= f_{C^*}(U_{C^*}) \\
W &= f_W(U_W) \\
S &= f_S(U_S) \\
H(k) &= f_{H(k)}(H(k-1), U_{H(k)}) \\
N(k) &= f_{N(k)}(W, N(k-1), H(k), A(k-1), U_{N(k)}) \\
A(k) &= f_{A(k)}(W, \bar{A}(k-1), N(k), U_{A(k)}) \\
Y^*(k) &= f_{Y^*(k)}(C^*, W, S, H(k), \bar{A}(k), U_{Y^*(k)}) \\
Y(k) &= f_{Y(k)}(C^*, Y^*(k))
\end{aligned}$$

for k from 1 to K, the maximum follow-up. The full cohort was generated according to $\mathcal{M}^F$. $U = \left(U_{C^*}, U_W, U_S, U_{H(k)}, U_{N(k)}, U_{A(k)}, U_{Y^*(k)}\right)$ denotes the set of random background factors that determine the values of

$$(C^*, W, S, H(k), N(k), A(k), Y^*(k), Y(k))$$

according to the deterministic functions

$$\left(f_{C^*}, f_W, f_S, f_{H(k)}, f_{N(k)}, f_{A(k)}, f_{Y^*(k)}, f_{Y(k)}\right).$$

Each equation reflects assumptions about how the data were generated by Nature [55]. Changing or intervening on one equation does not affect the remaining equations. The background factors U are assumed to be jointly independent in this particular model. A causal effect is defined in terms of the joint distribution of the observed data under an intervention on one or more of the structural equations.

3.2.3 Simulated data.

Data were simulated to reflect two aspects of the healthy-worker survivor effect, (i) time-varying confounding affected by prior exposure, and (ii) heterogeneity in susceptibility to exposure-related effects. Aside from time since hire to the start of the cancer registry, all covariates were Bernoulli ($\mathcal{B}$) random variables with a probability defined as a logit function of selected covariates. The full data was generated according to the following formulas:

Random Errors. $(U_{N(t)}, U_{E(t)}, U_{D(t)}, U_{Y(t)}) \sim Uniform[0,1]$

Baseline Covariates. $S = \mathcal{B}(p_S), W = \mathcal{B}(p_W), C^* = 0.$

Health Status. If $H(t-1) = 1$ then $H(t) = 1$. Otherwise, $H(t) \sim B(p_H)$.

At work. If $N(t-1) = 0$ then $N(t) = 0$. Otherwise,

$$N(t) \sim \mathcal{B}\left\{logit(\beta_0^N + \beta_W^N W + \beta_H^N H(t) + \left[\beta_E^N E(t-1)\right] I(t > 1) + U_{N(t)})\right\}.$$

Exposure. If $N(t) = 0$ then $E(t) = 0$. Otherwise,

$$E(t) \sim \mathcal{B}\left\{logit\left(\left[\beta_0^E + \beta_W^E W\right] I(t = 1) + \left[\beta_E^E E(t-1)\right] I(t > 1) + U_{E(t)}\right)\right\}.$$

Right Censoring. If $D(t-1) = 1$ then $D(t) = 1$. Otherwise,

$$D(t) \sim \mathcal{B}\left\{logit\left(\beta_0^D + \beta_W^D W + \beta_{\bar{E}}^D \bar{E}(t-1) + \left[\beta_Y^D \sum_{i=1}^{t-1} Y^*(i)\right] I(t > 1) + U_{D(t)}\right)\right\}.$$

Outcome. If $Y^*(t-1) = 1$ then $Y^*(t) = 1$. Otherwise,

$$Y^*(t) \sim \mathcal{B}\left\{logit\left(\beta_0^Y + \beta_W^Y W + \beta_E^Y E(t) + \left[\beta_{\bar{E}}^Y \bar{E}(t-1)\right] I(t > 1) + \beta_H^Y H(t) + \beta_S^Y S\bar{E}(t) + U_{Y(t)}\right)\right\}.$$

$$Y(t) = Y^*(t).$$

We considered five simulation scenarios in order to evaluate the impact of changing the parameters of the data generating distribution would have on the bias induced by left filtering. In the first set of simulations we present the base case (scenario 1), in which 10% of the workers are susceptible to exposure-related effects, the cumulative incidence of the outcome is comparable to the life-time risk of the most common cancers (approximately 12%), the log-odds ratio of mortality for each additional year since cancer diagnosis, β_Y^D, is 0.5, and moderate time-varying confounding affected by prior exposure is present. In scenario 2 we evaluate the effect of higher cancer-related mortality, by increasing the value of the coefficient β_Y^D. In scenario 3 we evaluate the impact of increasing the proportion of susceptible workers (p_S). In scenario 4 we evaluate how increasing the magnitude of time-varying confounding affects bias by increasing the effects of the adverse health status on leaving work (β_H^N), and cancer incidence (β_H^Y). In scenario 5 we increased disease incidence by increasing β_0^Y. The values of all coefficients for each set of simulations scenarios are presented in *Table 3.1.*

The observed data were generated as a function of the full data. Time from hire to the establishment of the cancer registry C^*, was generated as a uniformly distributed random variable ranging from 0 to 20. All indicators $Y(t)$ were set to 0 for those workers who developed an underlying cancer event ($Y^*(t) = 1, t \leq C^*$) prior the start of the registry, and the true cancer diagnoses indicators $Y^*(t)$ were dropped. All datasets were simulated using the *simcausal* R package.

3.2.4 Interventions, counterfactuals, and target parameters.

Under $\mathcal{M}^F$, each component of the data structure is generated as a deterministic function of its parents and the exogenous errors U. That is, we replace the intervention nodes $\bar{A}(k)$ in $f_{Y(k)}$ with those set by our rule, and previous $H(k), N(k)$ nodes by their previously generated counterfactual values. The counterfactual values of $Y(t)$ are then generated sequentially for each time point k. Following the Neyman-Rubin model, the counterfactual $Y_{\bar{d}}(k)$ is the outcome an individual would have at time k if, possibly contrary to fact, they had exposure assigned according to rule $\bar{d}$.

Since workers are occupationally exposed only while employed, we focus on the causal effects of dynamic treatment regimens, which assign exposure according to a worker's employment status. More formally, we define d_1 as an intervention that sets the right censoring node $D(k)$ to 0, and exposure node $E(k)$ to 1 while a worker is actively employed (while $N(k)$ is 1). However, once the value $N(k)$ changes to 0, d_1 then sets $E(k)$ to 0. Analogously, d_0 is a dynamic intervention that assigns workers to no exposure while employed and prevents right censoring. The counterfactual outcome $Y_{\bar{d}_1}(k)$ under exposure rule d_1 corresponds to the outcome that would have been observed if, possibly contrary to fact, a subject were always exposed while employed and remained uncensored. Similarly, the counterfactual outcome $Y_{\bar{d}_0}(k)$ corresponds to the outcome that would have been observed at time point k if, possi-

Coefficient	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_E^N	-1.50	-1.50	-1.50	-1.50	-1.50
β_0^E	-1.50	-1.50	-1.50	-1.50	-1.50
β_W^E	-0.50	-0.50	-0.50	-0.50	-0.50
β_E^E	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
$\beta_{\bar{E}}^D$	0.50	0.50	0.50	0.50	0.50
$\beta_{\bar{Y}}^D$	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_E^Y	0.25	0.25	0.25	0.25	0.25
$\beta_{\bar{E}}^Y$	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

Table 3.1: Coefficients of the data generating equations by scenario. For each scenario we indicate in bold the coefficient that differs from the base case (scenario 1).

bly contrary to fact, a subject were never exposed while at work and remained uncensored. The corresponding parameters for these interventions were the mean survival times, under each regimen. The exposure effect was measured by the difference between these quantities.

3.2.5 Estimating survival and evaluating bias.

Bias was measured as the difference between true exposure effects evaluated in the full data and the estimated exposure effects in the observed data. To determine the true exposure effects, for each scenario we generated a dataset of million workers followed for a maximum of 20 years according to the equations for the full data. We computed the counterfactual survival function of the full data under each of the regimens discussed in the previous section.

The survival function $S(t)$ expresses the probability that a person has not yet experienced the event of interest at the end of time point t, i.e. $P(T > t)$. In the previous section we state that we intervened on the structural equation model to generate the counterfactual outcomes $Y_{i,\bar{d}_1}(t), Y_{i,\bar{d}_0}(t)$ for all workers $i = 1, \ldots, n$ at time point $t = 1, \ldots, K$. The respective counterfactual survival curves are given by:

$$S^0_{\bar{d}_1}(t) = 1 - E\left(Y_{i,\bar{d}_1}(t)\right),$$

$$S^0_{\bar{d}_0}(t) = 1 - 1 - E\left(Y_{i,\bar{d}_0}(t)\right).$$

Expected time to a cancer diagnosis, or expected (mean) survival under each regimen, was computed as the area under the respective survival curve,

$$\mu^0_{\bar{d}} = \sum_0^K S^0_{\bar{d}}(t)dt.$$

Counterfactual exposure effect was measured in terms of the difference in the expected counterfatual survival of workers always exposed while at work and workers that were never exposed while work, given by:

$$\psi_0 = \mu^0_{\bar{d}_1} - \mu^0_{\bar{d}_0}.$$

Parameters of full-data counterfactuals represent the truth against which estimates are compared. In addition to a full dataset, for each scenario we simulated 500 observed datasets consisting of 50,000 workers. For each dataset, we estimated survival curves for workers following the regimens of interest using an adjusted Kaplan-Meier estimator [21], which adjusts for time-varying confounding affected by prior exposure, but ignores left filtering.

Since workers in the observed data cannot have an observed cancer diagnosis prior to their individual C^*, we used a delayed-entry adjusted Kaplan-Meier estimator to compute the survival in each of the simulated observed datasets [21, 22]. In addition to allowing workers to enter risk-sets at their C^*, the estimator adjusted for time-varying confounding affected by prior exposure and dependent right censoring [21, 22]. Details of the estimator and its implementation are provided in the next section.

3.2.6 The delayed-entry adjusted Kaplan-Meier estimator.

As in the previous section, let $S_{\bar{d}}(t)$ denote the survival function for workers following regimen $\left(\bar{d}_1, \bar{d}_0\right) \in \bar{d}$ among a group of n workers. Allowing ties, at time $t = 1, \ldots, K$ there are are $c_{\bar{d}}(t)$ incident cancer diagnoses out of $R_{\bar{d}}(t)$ workers at risk among regimen $\bar{d}$ followers. Since workers are at risk of having a reported incident cancer diagnosis only after the establishment of the registry,

$$c_{\bar{d}}(t) = \sum_{i}^{n} I\left(Y_i(t) = 1\right) \times I\left(\bar{A}_i(t) = \bar{d}\right) \times I\left(t \geq C_i^*\right),$$

where $I\left(Y_i(t) = 1\right)$ is an indicator that worker i is diagnosed with cancer at time t, $I\left(\bar{A}_i(t) = \bar{d}\right)$ is an indicator that worker i is following regiment $\bar{d}$ at time t, and $I\left(t \geq C_i^*\right)$ indicates that the registry was established before t. The number of workers at risk at t is given by

$$R_{\bar{d}}(t) = \sum_{i=1}^{n} I\left(\bar{A}_i(t) = \bar{d}\right) \times I\left(t \geq C_i^*\right).$$

Since cumulative exposure is associated with both the outcome of interest (incident cancers) and death (right censoring event), our estimator was adjusted for dependent censoring as well as non-random treatment assignment by weighting each subject at each time point by a weight that was inversely proportional to their probability of following rule $\bar{d}$ (i.e., being always or never exposed while at work and remaining uncensored). Inverse-probability estimators correct for informative censoring and confounded treatment assignment by giving extra weight to uncensored subjects who followed the regimen of interest. Details on the estimation of the weights are provided in the next section. Denoting $w_{i,\bar{d}}(t)$ as the weight of subject i at time point t_j, the weighted number of events and the weighted number at risk following regimen $\bar{d}$ are defined as

$$c_{\bar{d}}^{w}(t) = \sum_{i}^{n} w_{i,\bar{d}}(t) \times I\left(Y_i(t) = 1\right) \times I\left(\bar{A}_i(t) = \bar{d}\right) \times I\left(t \geq C_i^*\right),$$

and

$$R_{\bar{d}}^{w}(t) = \sum_{i=1}^{n} w_{i,\bar{d}}(t) \times I\left(\bar{A}_i(t) = \bar{d}\right) \times I\left(t \geq C_i^*\right).$$

The following formula defines the delayed-entry adjusted Kaplan-Meier estimator for followers of regimen $\bar{d}$:

$$\hat{S}_{\bar{d}}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{j \leq t}\left[1 - \frac{c_{\bar{d}}^{w}(j)}{R_{\bar{d}}^{w}(j)}\right] & \text{if } t_1 \leq t \end{cases}$$

if $R_{\bar{d}}^{w}(t) > 0$, and t_1 denotes the first failure time. Note that the current estimator differs from the adjusted Kaplan-Meier estimator proposed by Xie and Liu [14] in two ways. First, our equations for computing the number of incident events and subjects at risk incorporated an indicator that the registry was in place. Secondly, in addition to adjusting for confounding of the exposure-outcome relationship, we additionally adjusted for dependent censoring by incorporating the probability of remaining uncensored in the inverse probability weights. The estimated exposure effect is given by $\hat{\psi} = \hat{\mu}_{\bar{d}_1} - \hat{\mu}_{\bar{d}_0}$, where $\hat{\mu}_{\bar{d}}$ denotes the delayed-entry adjusted Kaplan-Meier estimated mean survival. Estimation of the survival functions for the observed data was carried out using the *stremr* R package [45].

3.2.7 Practical implementation of the inverse-probability weights

In order to estimate the weights $w_{i,\bar{d}}(t)$, we fit two logistic regression models at each time point:

$$\mathit{logit}\left[P\left(E(t)=1|W,D(t)=0,N(t)=1,E[t-1]\right)\right]=\alpha_0+\alpha_1 W+\alpha_2 E(t-1),$$

$$\mathit{logit}\left[P\left(D(t)=1|W,D(t-1)=0,\bar{E}(t-1)\right)\right]=\beta_0+\beta_1 W+\beta_2 \bar{E}(t-1).$$

The first model was fit among workers alive and at work at each time point, and was then used to estimate each employed worker's probability of exposure at t. We denote this probability as $\hat{p}_{i,e_1}(t)$. The predicted probability that a worker was always exposed at t was given by the cumulative product of the probabilities of being exposed at all time points up to t,

$$\hat{p}_{i,\bar{e}_1}(t)=\prod_{j=1}^{t}\hat{p}_{i,e_1}(j).$$

The predicted probability that a worker was always unexposed at t is:

$$\hat{p}_{i,\bar{e}_0}(t)=\prod_{j=1}^{t}\left(1-\hat{p}_{i,e_1}(j)\right).$$

The second model was used to estimate the conditional probability that a worker was censored during time point t, given that they were alive at the end of the previous time point ($D_i(t-1)=0$). We denote this probability as $\hat{p}_{i,c_1}(t)$. The probability that a worker remained *uncensored* up to t is given by

$$\hat{p}_{i,\bar{c}_0}(t)=\prod_{j=1}^{t}\left(1-\hat{p}_{i,c_1}(j)\right).$$

For each worker i and time point t, we computed the weights

$$w_{i,\bar{d}_1}(t)=\frac{1}{\hat{p}_{i,\bar{e}_1}(t)\times\hat{p}_{i,\bar{c}_0}(t)},$$

and

$$w_{i,\bar{d}_0}(t) = \frac{1}{\hat{p}_{i,\bar{e}_0}(t) \times \hat{p}_{i,\bar{c}_0}(t)}.$$

All analyses were performed in the R programming language [46].

3.3 Results

Plots of true and estimated survival curves for each of the 5 scenarios we considered are presented in *Figure 3.2*. The adjusted Kaplan-Meier estimated curves consistently overestimate the true (counterfactual) survival for both regimens and across all scenarios. This estimator controls for time-varying confounding affected by prior exposure, but ignores left filtering. Adjusting for time-varying confounding and taking into account left filtering, the delayed-entry Kaplan-Meier results in near unbiased estimates of exposure effects in four of the fives scenarios that we considered. Bias increased when disease incidence is increased to 25% in scenario 5.

In *Table 3.2* we present the subset of parameters of the data-generating distribution that vary across simulation scenarios. In addition, *Table 3.2* presents the true mean survival for workers following each of the regimens of interest, and exposure effects, evaluated in the full data. Exposure effects are expressed in terms of the number of years by which exposure reduced the mean survival. The mean estimated survival for each regimen and estimator, as well as the respective exposure effects, averaged across 500 simulated datasets of 50,000 workers, are also provided in *Table 3.2.* Our measure of bias indicates the number of years by which exposure effect estimates under- or overestimate the true effect of exposure. For example, the adjusted Kaplan-Meier exposure effect underestimates the true harmful effects of exposure by more than half a year (0.56 years) in scenario 1. However, a bias measure of -0.04 indicates that the delayed-entry Kaplan-Meier exposure effect overestimates the true harmful effect of exposure by approximately 0.04 years (15 days) over the 20-year follow-up. The delayed-entry Kaplan Meier estimates had greater variability than the adjusted Kaplan-Meier estimates across all scenarios.

Scenario 1 (the base case) presents a setting in which 12% of the worker population were susceptible to exposure-related effects on incident cancers, and the 20-year cumulative incidence of the outcome under study is approximately 12%. Bias remained practically unchanged (-0.05 years in scenario 2) when cancer-related mortality was increased, or when the proportion of susceptible workers was doubled (-0.04 years in scenario 3). We evaluated the impact of the confounding effect of health status by simultaneously increasing the effects of poor health on leaving work and on the outcome of interest (scenario 4); bias increased slightly, from -0.06 years in the base case to -0.12 years in this scenario. Lastly, we evaluated the effect of increasing the incidence of the outcome (scenario 5); bias was largest under this scenario (-0.21 years).

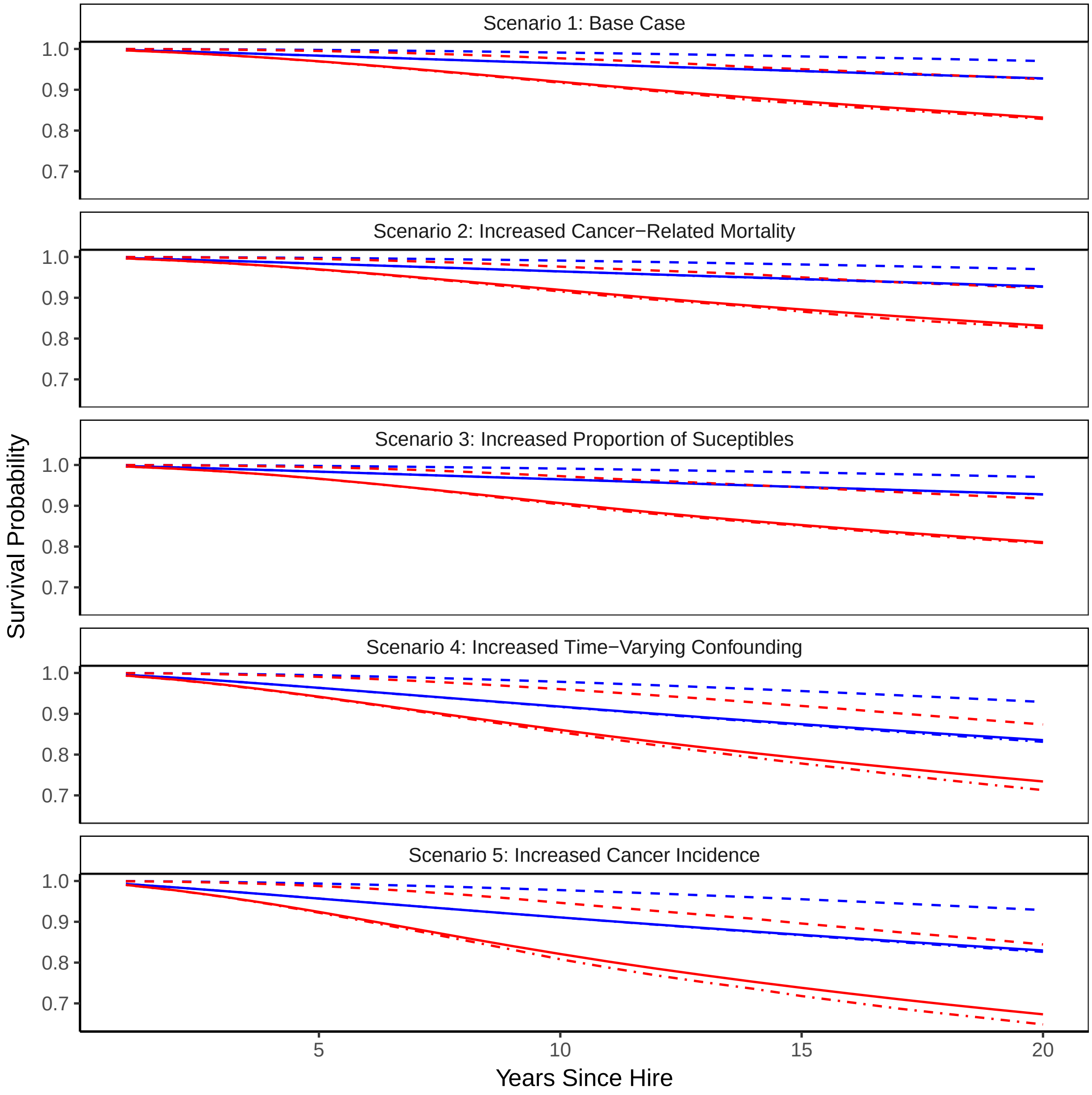


Figure 3.2: True survival curves (*solid*), and estimated adjusted Kaplan-Meier (*dashed*), and delayed-entry adjusted Kaplan-Meier (*dash-dot*) survival curves among cohorts of workers that are never exposed (*blue*) and always exposed (*red*) while at work. Scenario 1 represents the base case. Scenario 2 evaluates the role of higher cancer-related mortality. In scenario 3 the proportion of susceptible workers is increased. The magnitude of time-varying confounding affected by prior exposure is increased in scenario 4. Cancer incidence is increased in scenario 5.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
Parameters of the data generating distribution					
Proportion of susceptible workers, pS	10	10	20	10	10
Coefficient for year since diagnosis on mortality, $\beta_{\bar{Y}}^{D}$	0.5	2.0	0.5	0.5	0.5
Coefficient for health status on the disease, β_H^Y	0.7	0.7	0.7	0.7	0.7
Coefficient for health status on leaving work, β_H^N	0.5	0.5	0.5	0.5	0.5
Cumulative incidence in the full data	12.0	12.0	13.0	21.5	25.0
Simulation results in years					
The truth					
Mean survival if always exposed while at work	17.39	17.39	17.17	16.31	15.61
Mean survival if never exposed while at work	18.29	18.29	18.29	17.38	17.26
Difference in mean survival (exposure effect)	-0.90	-0.90	-1.13	-1.06	-1.65
Observed (right censored and left filtered) data, mean (SE)					
Adjusted Kaplan-Meier					
Mean survival if always exposed while at work	18.44 (0.16)	18.43 (0.22)	18.36 (0.16)	18.06 (0.09)	17.79 (0.28)
Mean survival if never exposed while at work	18.78 (0.01)	18.78 (0.01)	18.78 (0.01)	18.48 (0.01)	18.47 (0.02)
Difference in mean survival (exposure effect)	-0.34 (0.16)	-0.35 (0.22)	-0.42 (0.24)	-0.42 (0.09)	-0.68 (0.28)
Bias	0.56	0.55	0.71	0.64	0.97
Delayed-Entry Adjusted Kaplan-Meier					
Mean survival if always exposed while at work	17.35 (0.31)	17.33 (0.35)	17.13 (0.40)	16.17 (0.18)	15.35 (0.45)
Mean survival if never exposed while at work	18.29 (0.03)	18.29 (0.04)	18.29 (0.03)	17.35 (0.05)	17.24 (0.06)
Difference in mean survival (exposure effect)	-0.94 (0.31)	-0.96 (0.35)	-1.16 (0.40)	-1.18 (0.19)	-1.86 (0.45)
Bias	-0.04	-0.05	-0.04	-0.12	-0.21

Table 3.2: Selected simulation parameters and simulation results by scenario.

3.4 Discussion

Although worker exposures to MWFs in the UAW-GM cohort started in the 1940s, cancer incidence prior to the establishment of the Michigan Cancer Registry in 1985 is unknown. If the cancer incidence sub-cohort of workers alive in 1985 consists of workers who were less susceptible to the adverse effects of MWF exposure, then it is a biased representation of the full cohort. In this first formal evaluation of the bias introduced by left filtering, we found that estimators that ignore left filtering lead to effect estimates that are biased downward, and the magnitude of the bias increases with increasing incidence of the disease under study, or with increasing proportion of susceptible workers. However, addressing left-filtering with the proposed delayed-entry Kaplan-Meier estimator results in little bias when controlling for time-varying confounding affected by prior exposure, at least when disease incidence is not too high. In addition, the magnitude of bias in the latter estimator was not affected by increases in the proportion of susceptible workers, or the magnitude of time-varying confounding. However, bias increased when the disease incidence was doubled.

In a previous simulation study, Applebaum et al. evaluated the bias in hazard ratios for mortality estimated in left truncated occupational cohorts in the presence of heterogeneity in susceptibility [15]. Those analyses did not adjust for left truncation and thus were more analogous to our adjusted Kaplan-Meier estimates. Consistent with our current findings, authors reported downward bias in exposure effect estimates in left truncated cohorts when susceptibility with respect to exposure-related health effects varied between workers [15]. This study extends that work in several ways. We evaluated bias in the presence of time-varying confounding affected by prior exposure, as well as heterogeneous susceptibility. Attentive to Hernàn's caution against reporting a summary hazard ratio when the depletion of susceptible results in decreasing hazard ratios over time [63], we reported full survival curves in addition to summary measures. In addition, we have proposed an estimator (the delayed-entry Kaplan-Meier estimator) that leads to unbiased estimates in most common simulation scenarios, when the time of hire is assumed random (i.e., does not depend on measured or unmeasured factors).

The depletion of susceptible subjects has also been noted as a source of selection bias in several non-occupational applications. Selection bias explains the apparently disparate results from studies of cigarette smoking in relation to dementia risk, where smoking is harmful overall but appears beneficial at older ages [64]. Most smokers who are susceptible to developing dementia due to their smoking do so at earlier ages, and thus older groups without dementia at baseline are depleted of susceptible smokers [64]. Similarly, a recent study of delayed lead exposure assessment in the Normative Aging Study reported increasing hazard ratios after adjustment for loss to follow-up between initial enrollment and the bone-lead sub-study [65]. Finally, birth cohorts are also affected in a similar manner since eligibility is typically based on live births. Exposure–outcome effect estimates can be biased if the exposure is associated with reduced likelihood of conception or increased fetal loss [66].

Our work provides further evidence that, if inappropriately controlled, selective inclusion in sub-studies of environmental and occupational exposures can substantially bias results.

Aside from left filtering, another feature of the UAW-GM cohort is that follow-up continues past employment termination. As unemployed workers cannot be exposed, the data structure results in a violation of the positivity assumption. This has been used to justify the application of G-estimation of accelerated failure time models in occupational epidemiology [53, 54], as the causal interpretation of these parameters does not rely on the positivity assumption. For rare outcomes such as cancer, however, there are alternative parameters of realistic interventions that also avoid positivity violations in settings where follow-up extends past employment termination. For example, a dynamic regimen that assigns exposure levels in response to subject covariate values [55] makes it possible to define interventions under which a worker is assigned nonzero exposures only while at work. In this study, we evaluated causal parameters of dynamic interventions in a setting where worker follow-up continued past employment termination. The effects of such interventions have been previously assessed in the field of occupational epidemiology in few simulated or applied examples [56, 57, 67]. A possible application of these workplace interventions is the estimation of the risk reduction due to lowering occupational exposure limits below a priori determined cutoffs. A number of methods have been developed to estimate effects of dynamic intervention from observational data. One such method, targeted maximum likelihood estimation (TMLE) [37], uses the efficient estimating equation framework [38, 40] to produce estimators that adjust for time-varying confounding, are consistent under partial model misspecification (double robustness property), and are efficient when models for treatment mechanism and outcome are correctly specified (local efficiency) [37]. However, to our knowledge none of these approaches have been extended to address left filtering.

In conclusion, this simulation study has demonstrated, for the first time, that left filtering in the presence of the healthy worker survivor effect induces bias. The delayed-entry adjusted Kaplan-Meier estimator, proposed here, is one analytic approach that appears to reduce the bias.

Chapter 4

Etiologic effect measures versus causal estimates of real world workplace interventions

4.1 Introduction

The risk of occupational hazards is often estimated from observational studies in which exposures are not randomly assigned, but may depend on a number of factors [1–5, 10]. Intermediate health status, for example, may be predicted by past exposure, and may also predict future exposure and outcomes. Time-varying confounding by factors in the causal pathway between past exposure and the outcome is a key aspect of the Healthy Worker Survivor Effect (HWSE) in observational cohort studies of occupational disease. The HWSE is a ubiquitous process that results in the healthiest workers accruing the most lifetime exposure while less healthy workers limit their exposure by reducing the time at work, switching to lower exposed jobs, or altogether leaving the workforce. Driven by the HWSE, studies may estimate null, or even protective effects of occupational hazards.

The potential outcomes framework defines causal effects as contrasts of the distributions of counterfactual (potential) outcomes under hypothetical interventions that assign exposure levels and may further prevent censoring [49]. Interventions that deterministically assign the same exposure and/or censoring values to all subjects in a population are known as static interventions [51]. For example, the causal relationship between airborne particulate matter ($PM_{2.5}$) and heart disease was inferred by contrasting the 15-year cumulative incidence of heart disease under an intervention that exposed a cohort of active aluminum workers above the median $PM_{2.5}$, to the 15-year cumulative incidence that would have been observed under a second intervention that exposed the same cohort below the median $PM_{2.5}$ [68]. Both interventions prevented leaving work before the age of 55, the normal age of retirement [68]. Exposure and employment status were set rather than predicted by the past under these interventions. Workers, who might transfer to jobs with more or less exposure or terminate employment as a function of their health status, were instead forced to remain at work and receive their assigned exposure for the entire duration of the study. As a result the exposure assignment and employment termination processes were not confounded under these interventions. We refer to such causal contrasts that are not affected by the HWSE as etiologic effects, since they reflect the exposure-outcome relationship in the absence of the influence of factors that determine which workers get exposed.

Counterfactuals can additionally be defined under dynamic interventions that assign exposure and/or censoring as a function of one's observed past, in contrast to assigning the same value(s) to all subjects in a population. For example, a study of diesel exhaust and lung cancer compared the risk under a series of interventions that set occupational exposure to hypothetical levels while workers were actively employed [56]. While exposure levels under each intervention were determined *a priori*, the interventions permitted health-related early employment termination. We refer to the latter as realistic interventions, because they reflect the real-world self-selection of workers out of the workforce. Exposure-response under realistic interventions is affected by the HWSE. Several recent occupational applications have used the parametric g-formula to evaluate parameters of realistic interventions [56, 57, 67]. In contrast to etiologic interventions that aim to estimate the biologic effect of long-term, sustained exposure, realistic interventions aim to estimate the expected disease experience of a population under a proposed standard.

Exposure effect estimates of the two interventions can differ substantially. Consider a workplace study where workers exposed to a high level of an occupational hazard may leave work earlier (accumulating less exposure) than they would have if exposed at a low level. Etiologic interventions would always result in higher risk estimates under the high exposure scenario in this setting. It is possible that in the same study, realistic interventions may results in lower risk estimates for the population under the lower exposure scenario. Ideally, the choice of intervention whose effects are estimated would be motivated by the research question. In practice, the choice is often dictated by the available data. For example, in occupational studies where follow-up continues past employment termination, unemployed workers have a zero probability of exposure, resulting in violations or near violations of the positivity assumption, also known as experimental treatment assignment. The assumption posits sufficient variability in exposure assignment within strata of confounders, and is required for identifiability of causal effects [6, 50, 60]. Etiologic effects may be non-identifiable in workplace settings where follow-up extends past employment. In contrast, a well posed realistic intervention may avoid positivity violations resulting in identifiable causal effects.

Moreover, the National Institute for Occupational Safety and Health (NIOSH) recommended exposure limits, and Occupational Safety and Health Administration (OSHA) permissible exposure limits are based [69], that correspond to etiologic interventions. When the estimation of such effects is infeasible, estimates of realistic interventions may be used to assess risk. It is therefore important to understand under what conditions the two effect measures are comparable, and to identify factors that drive the differences between the two. In this simulation study we evaluated how the relationship between the two exposure-response measures varies as a function of key drivers of the HWSE, namely: (i) the strength of the relationship between intermediate health status and outcome, and the temporal relationship between (ii) intermediate health status and leaving work, and (iii) exposure and health status relationship.

4.2 Methods

4.2.1 Data description and notation

We denote the underlying data structure of every participant in an occupational cohort as a realization of the random variable X. For each worker and year t of follow-up, we simulated a set of time-dependent covariates that included $W(t)$: an indicator of active employment status in year t; $E(t)$: an indicator of exposure to the occupational hazard under study in year t; $Y(t)$: an the indicator of a diagnosis with the outcome of interest, such as lung cancer, on or before year t; and $H(t)$: an indicator of poor health, such as a diagnosis with an adverse health event or an underlying chronic condition. A bar over a variable denotes the history from baseline. For example, $\bar{E}(t)$ includes the entire exposure history from year 1 to year t. In addition to time-dependent covariates, we simulated S, an indicator of a worker's genetic susceptibility to adverse health outcomes $H(t)$ and $Y(t)$. In summary, the data on n independent and identically distributed (iid) subjects was defined as:

$$X_i(t) = (S, \bar{W}_i(t), \bar{E}_i(t), \bar{H}_i(t), \bar{Y}_i(t))$$

for $i = 1, \ldots, n$.

4.2.2 Simulated data-generating distributions

We simulated data under five hypothetical scenarios. All covariates were generated as uniform random variables ranging between 0 and 1, or as Bernoulli ($\mathcal{B}$) random variables with probabilities defined as logit-linear functions of selected covariates and an error term. Only susceptible workers were at risk of experiencing $H(t)$ and/or $Y(t)$ in all scenarios. In addition we assumed that workers who terminated employment could not be hired at a later date, and that once diagnosed with the adverse health event H(t), workers experienced it for the remainder of follow-up.

The first three scenarios explore how the difference between etiologic and realistic exposure effect measures changes as a function of the strength of the association between mediator ($H(t)$) and leaving work ($W(t)$), one of the key relationships that drives the HWSE. In the

base case (Scenario 1), health status has a moderate effect on the probability of leaving work (coefficient $\beta_H^Y = 0.50$ in the equations below). The data for each worker in this scenario was generated according to the following longitudinal structural equation model [48] for $t = 1, \ldots, 20$ years of follow-up:

Random Errors. $(U_{E(t)}, U_{H(t)}, U_{Y(t)}) \sim Uniform[-1, 1]$

Susceptibility. $S \sim \mathcal{B}(0.50)$

Exposure. $E(t) = \mathit{logit}\left(-1.00 + 2.00 \times E(t-1) + U_{E(t)}\right)$.

Health status. If $H(t-1) = 1$ then $H(t) = 1$. Otherwise,

$$H(t) = \mathit{logit}\left(-2.00 + 0.25 \times \bar{E}(t) + U_{H(t)}\right) \times I(S = 1).$$

Employment status. $W(t) = 0$ if $H(t-1) = 1$. Otherwise $W(t) = 1$.

Outcome.

$$\begin{aligned} Y(t) &= \mathit{logit}\left(\beta_0^Y + \beta_H^Y \times H(t) + \beta_E^Y \times \bar{E}(t) + U_{Y(t)}\right) \times I(S = 1). \\ &= \mathit{logit}\left(-7.00 + 0.50 \times H(t) + 0.40 \times \bar{E}(t) + U_{Y(t)}\right) \times I(S = 1). \end{aligned}$$

In **Scenario 2** we increased the effect of the mediator on the outcome ($\beta_H^Y = 3.00$), changing the equation for the outcome to the following:

$$Y(t) = \mathit{logit}\left(-7.00 + 3.00 \times H(t) + 0.40 \times \bar{E}(t) + U_{Y(t)}\right) \times I(S = 1).$$

The remaining equations were the same for this scenario as for the base case. The causal relationships between the variables in the first two scenarios are also presented in the directed acyclic graph (DAG) [47, 48] in *Figure 4.1.* This representation indicates that the mediator $H(1)$ is both a confounder of the $E(2) \rightarrow Y(2)$ relationship in the $E(2) \leftarrow W(2) \leftarrow H(1) \rightarrow H(2) \rightarrow Y(2)$ pathway, and a mediator of the $E(1) \rightarrow Y(1)$ relationship in the$E(1) \rightarrow H(1) \rightarrow Y(1)$ pathway. Susceptibility is a modifier of the effect of the exposure on these two factors.

In **Scenario 3** the adverse health event does not increase the risk of the outcome and therefore it is not a mediator of the past exposure-outcome relationship. For example, occupational silica exposure is known to increase the risk of both silicosis ($H(t)$), and lung cancer ($Y(t)$) [70]. However, it is unclear whether silicosis itself is a risk factor for lung cancer [70]. The lack of a causal relationship between the two is reflected by the absence of an arrow between $H(t)$ and $Y(t)$ in *Figure 4.2.*

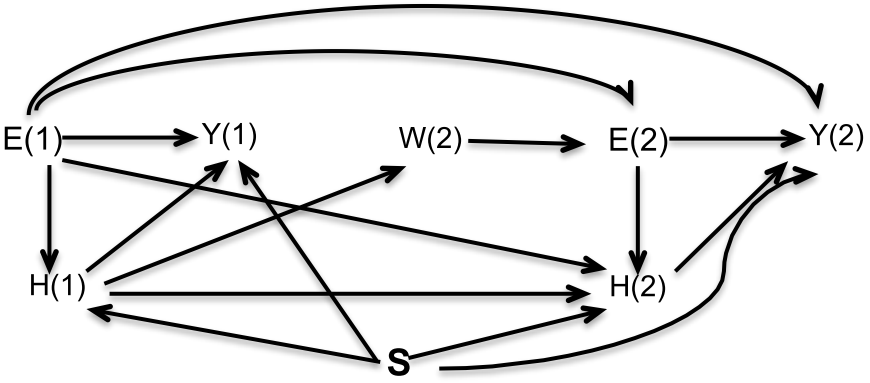


Figure 4.1: DAG for scenarios 1 and 2. Health status ($H(t)$) and the outcome ($Y(t)$) are both predicted by cumulative exposure.

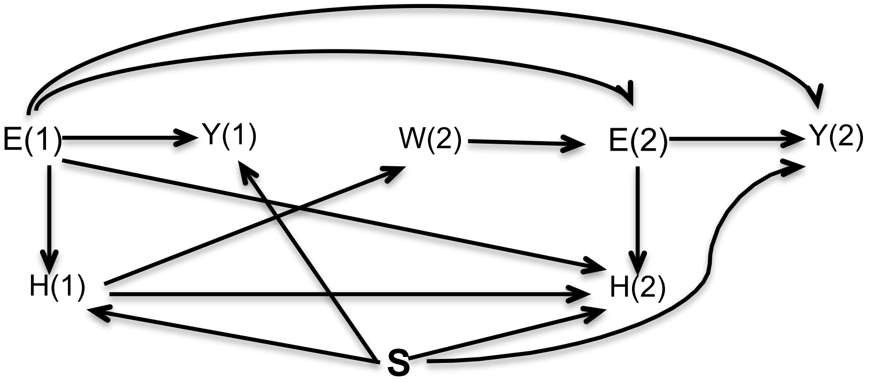


Figure 4.2: DAG for scenario 3. Health status does not predict the outcome.

While all other covariates were generated as in the base case, the outcome in this scenario was generated according to the following equation:

$$Y(t) = logit\left(-7.00 + 0.00 \times H(t) + 0.40 \times \bar{E}(t) + U_{Y(t)}\right) \times I(S = 1).$$

Scenario 4 evaluates how the temporal relationship between the mediator and leaving work impacts etiologic and realistic exposure effect measures. For example, while diabetes ($H(t)$) increases the risk of cardiovascular events ($Y(t)$), a number of years may pass between first diagnosis of diabetes and the appearance of advanced diabetes symptoms that may force workers to terminate employment. To reflect this temporal relationship, active employment status in this scenario is predicted by adverse health events occurring 10 years prior, as follows:

If $H(t - 10) = 1$ or $W(t - 1) = 0$ then $W(t) = 0$. Otherwise $W(t) = 1$.

All other covariates were generated as in the base case. Causal relationships for this scenario are presented in *Figure 4.3*.

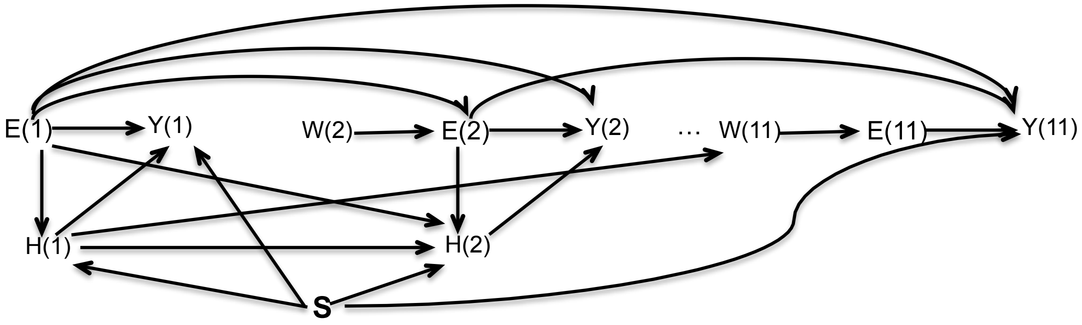


Figure 4.3: DAG for scenario 4. Active employment status is predicted by health status 10 years prior ($(H(t-10) \rightarrow W(t))$).

Lastly, in **Scenario 5** we evaluated the role of the timing of the exposure-mediator relationship. While cumulative exposure predicted health status and outcome in scenarios 1 through 4, here we consider an extreme setting in which only exposures in the first year of employment predict health status. Consider susceptible workers exposed to a skin irritant that ultimately causes melanoma. They may immediately experience an acute inflammatory response that may itself increase melanoma risk many yars later (*Figure 4.4*). If all susceptible subjects develop the acute condition right away and leave work, this is a "perfect screen" for preventing melanoma.

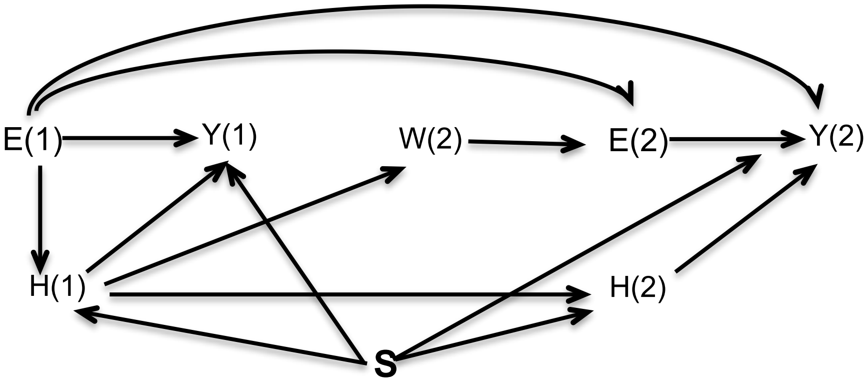


Figure 4.4: DAG for scenario 5. Health ($(H(t))$) is predicted by exposure in the first year of employment. The outcome $Y(t)$ is affected by cumulative exposure.

Health status in this scenario was generated as follows:

$$H(1) = \left(-4.00 + 7.00 \times E(1) + U_{H(1)}\right), \; H(t > 1) = H(1).$$

Remaining covariates were generated as in the base case.

4.2.3 Interventions and counterfactuals

As previously noted, our realistic interventions are dynamic interventions that assign exposure according to employment status. Specifically, we define $d_{r,1}$ as an intervention that sets the exposure node $E(t)$ to 1 while a worker is actively employed ($W(t) = 1$). However, once the worker terminates employment ($W(t) = 0$), $d_{r,1}$ sets $E(t)$ to 0. The counterfactual outcome $Y_{i,\bar{d}_{r,1}}(t)$ corresponds to the outcome that worker i would have in year t if they were always exposed while at work. Intervention $d_{r,0}$ assigns workers to no exposure before and after employment termination. We denote $Y_{i,\bar{d}_{r,0}}(t)$ as the counterfactual outcome for worker i at year t under the latter intervention.

Our etiologic interventions $\{d_{e,1},\ d_{e,0}\}$ are static interventions that set binary exposure $E(t)$ to either 1 or 0, and $W(t)$ to 1 for all years. We denote $Y_{i,\bar{d}_{e,1}}(t)$ and $Y_{i,\bar{d}_{e,0}}(t)$ as the counterfactual outcomes that worker i would experience in year t if they were always at work, and respectively always exposed or always unexposed. Counterfactual outcomes under each intervention were generated by setting nodes $E(t), W(t)$ in the system of equations above as specified by our interventions, sequentially for each subject $i = 1, \ldots, n$ and year of follow-up $t = 1, \ldots, 20$.

4.2.4 Exposure Effects

Denoting $\bar{d} \in \left(\bar{d}_{r,1}, \bar{d}_{r,0}, \bar{d}_{e,1}, \bar{d}_{e,0}\right)$ as one of the four regimens of interest, we computed counterfactual survival curve for each regimen in $\bar{d}$. The survival function $S(t)$ expresses the probability that a worker has not yet experienced the outcome of interest by the end of year t:

$$S^0_{\bar{d}}(t) = 1 - E\left(Y_{i,\bar{d}}(t)\right).$$

Expected time to the outcome, or expected (mean) survival under each regimen $\bar{d}$ was computed as the area under the respective survival curve,

$$\mu^0_{\bar{d}} = \int_0^K S^0_{\bar{d}}(t)dt.$$

The realistic exposure effect $\psi^0_r = \mu^0_{\bar{d}_{r,1}} - \mu^0_{\bar{d}_{r,0}}$, contrasts the mean survival of a cohort that is always exposed while at work to the mean survival if the same cohort were never exposed while at work: The etiologic exposure effect $\psi^0_e = \mu^0_{\bar{d}_{e,1}} - \mu^0_{\bar{d}_{e,0}}$, measures the difference in the mean survival of a cohort that is always at work and exposed, to the mean survival if the same cohort is always at work but unexposed. In order to contrast the effects of the two intervention classes, we computed the ratio of the etiologic and realistic exposure effect measures.

For each simulation scenario we generated 200 datasets of$n = 50,000$ workers. Survival, exposure effect estimates, and their ratio for each year of follow-up were averaged across the datasets.

All datasets were simulated using the *simcausal* R package [71]. All analyses were performed in the R programming language [46].

4.3 Results

In *Table 4.1* we present the distribution of all covariates for each scenario and year of follow-up, among workers following the realistic intervention $d_{r,1}$ under which they are always exposed while at work. In the last year of follow-up ($t = 20$), the proportion of susceptible workers is smaller when health status strongly predicts the outcome in scenario 2, than in other scenarios. The proportion of actively employed workers and cumulative exposure is also higher in scenario 2 than in the base case, since survivors who are less likely to be susceptible to the effects of exposure are more likely to remain at work. Cumulative exposure at the end of follow-up is greatest in scenario 4 where the relationship between health status and leaving work is lagged by 10 years.

In *Figures 4.5-4.7* we present comparisons of two or more simulation scenarios, each intended to illustrate how the survival experience of worker cohorts varies as a function of a specific aspect of the data generating process under the two interventions. For each scenario we present survival curves for cohorts of workers that were always exposed and never exposed according to etiologic and realistic interventions, respectively. For example, *Figure 4.5a* indicates that survival was identical for never-exposed workers under both the realistic and etiologic interventions, shown by overlaid survival curves. However always-exposed workers experienced a worse survival under the etiologic intervention. While survival at the end of follow-up was approximately one for workers always exposed under the realistic intervention, survival was less than 0.80 for workers always exposed under the etiologic intervention (*Figure 4.5a*). The exposure effect for each intervention was computed as the difference between the area under the always exposed survival curve and the area under the never exposed survival curve.

Along with etiologic and realistic exposure effect measures for each scenario, in *Table 4.2* we report the etiologic-to-realistic effect ratio, over time. Etiologic effects were greater than the realistic effects in all scenarios since we simulated harmful exposures that shorten survival. For example, an etiologic effect of 1.35 years ($t = 20$) for the base case indicates that, if at work and exposed for the duration of follow-up, workers would experience the event 1.35 years sooner than if they had been at work but unexposed for the same time period. A realistic exposure effect of 0.26 years in the same scenario indicates that workers who were exposed while at work, but could terminate employment for health-related reasons, experienced the outcome approximately 3 months sooner than if they had been unexposed.

An etiologic-to-realistic effect ratio of 4.13 in year 20 for scenario 1, denoted by $R_1(20)$, indicates that by leaving work as a result of exposure-related poor health, workers experienced a four-fold increase in their mean survival.

The first three simulation scenarios explore how the etiologic-to-realistic effects ratio changes as a function of the strength of the association between health status and leaving work (*Figure 4.5*). Etiologic and realistic exposure effects were most alike in scenario 2 where health status had a strong effect on the outcome (*Table 4.2*). The similarity between the two effect measures in scenario 2 is also reflected by smaller ratios in this scenario than in scenarios 1 and 3 (*Table 4.2*). The two effects were most dissimilar in scenario 3 where the outcome was not affected by health status, as reflected by the greater etiologic-to-realistic effect ratios ($R_1(20) = 4.13, R_2(20) = 2.16, R_3(20) = 6.86$).

In the second comparison (*Figure 4.6*) we contrast the base case with scenario 4 to explore how the ratio changes as a function of the temporal relationship between intermediate health status and leaving work. While health status in the previous year predicts leaving work in the base case, in scenario 4 leaving work is predicted by health events occurring 10 years earlier. Etiologic exposure effects were the similar in both scenarios (2.61 vs. 2.53 in year 20). However, realistic exposure effects in scenario 4 were greater than in the base case (0.63 vs. 2.21 in year 20), approximating the etiologic exposure effect. Consequently ratios were greater in the base case ($R_1(20) = 4.13, R_4(20) = 1.15$).

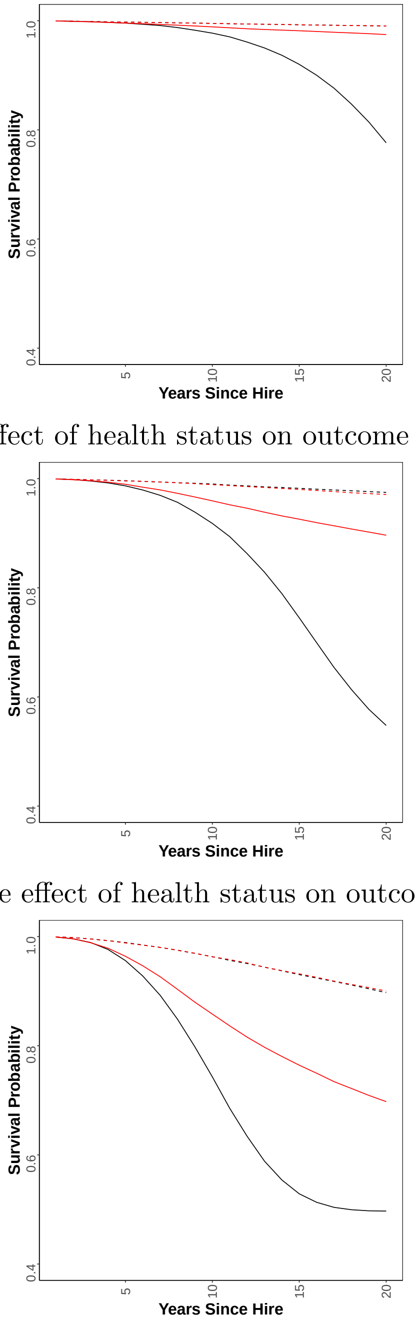
The last comparison (*Figure 4.7*) contrasts the base case to scenario 5, the latter representing a setting where workers experience the adverse health event during their very first year of exposure. Health status in this setting is a perfect screen for preventing the outcome. While etiologic effects were similar in both scenarios, realistic effects were smaller in scenario 5 (0.63 vs. 0.43 years). The etiologic-to-realistic effect ratios were larger in scenario 5 than in the base case ($R_1(20) = 4.13, R_5(20) = 9.86$).

4.4 Discussion

We evaluated the relationship between causal effects of etiologic workplace interventions that assign exposure and prevent leaving work, and causal effects of realistic interventions that assign exposure while workers are actively employed, allowing them to terminate employment. We found that effects of interventions that require workers to remain at work and receive their assigned exposure were always greater. The etiologic-to-realistic effects ratio represents the reduction (in the multiplicative scale) in exposure-related risk as a result of limiting exposure by early employment termination. The ratio decreased with increasing strength of the health status outcome relationship, and with synchronicity of the health-status and employment termination. The ratio increased when workers terminated employment early for exposure-related health reasons.

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$	$t=10$	$t=11$	$t=12$	$t=13$	$t=14$	$t=15$	$t=16$	$t=17$	$t=18$	$t=19$	$t=20$
Scenario 1																				
$E[S]$	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.48	0.48	0.48	0.48	0.48
$E[W(t)]$	1.00	1.00	0.91	0.82	0.73	0.66	0.60	0.56	0.53	0.52	0.51	0.51	0.51	0.51	0.51	0.52	0.52	0.52	0.52	0.52
$E[E(t)]$	1.00	1.00	0.91	0.82	0.73	0.66	0.60	0.56	0.53	0.52	0.51	0.51	0.51	0.51	0.51	0.52	0.52	0.52	0.52	0.52
$E[\bar{E}(t)]$	1.00	2.00	2.90	3.72	4.45	5.11	5.71	6.27	6.81	7.33	7.85	8.37	8.89	9.41	9.93	10.46	10.99	11.52	12.05	12.59
$E[H(t)]$	0.00	0.09	0.18	0.27	0.34	0.40	0.44	0.47	0.48	0.49	0.49	0.49	0.49	0.49	0.49	0.48	0.48	0.48	0.48	0.48
$E[Y(t)]$	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
Scenario 2																				
$E[S]$	0.50	0.50	0.50	0.49	0.49	0.48	0.47	0.46	0.45	0.43	0.42	0.40	0.39	0.37	0.36	0.34	0.33	0.32	0.31	0.30
$E[W(t)]$	1.00	1.00	0.91	0.82	0.74	0.68	0.63	0.60	0.58	0.58	0.59	0.60	0.61	0.63	0.64	0.66	0.67	0.68	0.69	0.70
$E[E(t)]$	1.00	1.00	0.91	0.82	0.74	0.68	0.63	0.60	0.58	0.58	0.59	0.60	0.61	0.63	0.64	0.66	0.67	0.68	0.69	0.70
$E[\bar{E}(t)]$	1.00	2.00	2.91	3.73	4.48	5.17	5.82	6.44	7.05	7.67	8.32	8.98	9.68	10.40	11.15	11.93	12.72	13.54	14.38	15.24
$E[H(t)]$	0.00	0.09	0.18	0.26	0.33	0.38	0.41	0.43	0.43	0.43	0.41	0.40	0.39	0.37	0.36	0.34	0.33	0.32	0.31	0.30
$E[Y(t)]$	0.001	0.003	0.007	0.011	0.015	0.019	0.022	0.025	0.026	0.026	0.025	0.024	0.023	0.022	0.020	0.019	0.018	0.017	0.016	0.015
Scenario 3																				
$E[S]$	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49
$E[W(t)]$	1.00	0.93	0.85	0.77	0.70	0.64	0.59	0.55	0.53	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
$E[E(t)]$	1.00	0.93	0.85	0.77	0.70	0.64	0.59	0.55	0.53	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
$E[\bar{E}(t)]$	1.00	1.93	2.77	3.55	4.24	4.88	5.47	6.02	6.55	7.06	7.57	8.08	8.60	9.11	9.62	10.14	10.65	11.17	11.69	12.21
$E[H(t)]$	0.07	0.15	0.23	0.30	0.36	0.41	0.45	0.47	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49
$E[Y(t)]$	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Scenario 4																				
$E[S]$	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.49	0.49	0.49	0.48	0.48	0.47	0.46	0.45	0.44	0.43	0.42	0.41	0.40
$E[W(t)]$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.91	0.83	0.75	0.69	0.64	0.61	0.60	0.60	0.61
$E[E(t)]$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.91	0.83	0.75	0.69	0.64	0.61	0.60	0.60	0.61
$E[\bar{E}(t)]$	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.00	11.00	11.91	12.74	13.49	14.18	14.83	15.46	16.09	16.72	17.37
$E[H(t)]$	0.00	0.09	0.18	0.27	0.34	0.40	0.44	0.47	0.48	0.48	0.48	0.48	0.47	0.46	0.45	0.44	0.43	0.42	0.41	0.40
$E[Y(t)]$	0.001	0.001	0.001	0.002	0.002	0.003	0.004	0.005	0.007	0.009	0.011	0.013	0.016	0.017	0.019	0.020	0.020	0.020	0.019	0.018
Scenario 5																				
$E[S]$	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49
$E[W(t)]$	1.00	0.52	0.52	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53
$E[E(t)]$	1.00	0.52	0.52	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53
$E[\bar{E}(t)]$	1.00	1.52	2.05	2.58	3.10	3.63	4.16	4.69	5.22	5.75	6.28	6.81	7.35	7.88	8.41	8.94	9.47	10.00	10.52	11.05
$E[H(t)]$	0.48	0.48	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47
$E[Y(t)]$	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.003	0.003

Table 4.1: Distribution of Simulated Covariates Among Workers Always Exposed While Active, by Scenario.

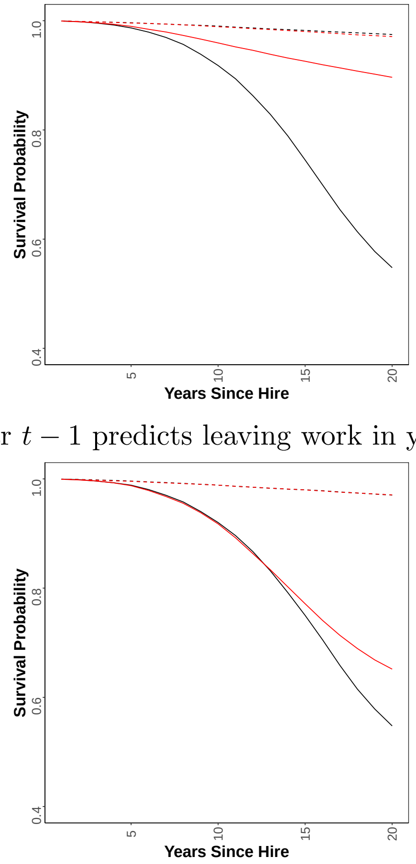


(a) No effect of health status on outcome ($\beta_H^Y = 0$)

(b) Moderate effect of health status on outcome ($\beta_H^Y = 1$)

(c) Strong effect of health status on outcome ($\beta_H^Y = 3$)

Figure 4.5: Evaluating the role of the health status - outcome relationship. Counterfactual survival curves among cohorts of exposed (solid) and unexposed (dashed) workers following etiologic (black) and realistic (red) interventions.

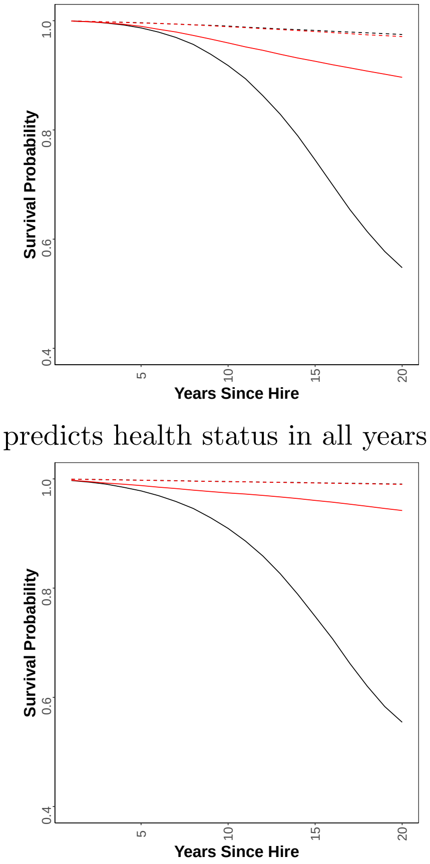


(a) Health status in year $t-1$ predicts leaving work in year t ($H(t-1) \rightarrow W(t)$)

(b) Health status in year $t-10$ predicts leaving work in year t ($H(t-10) \rightarrow W(t)$)

Figure 4.6: Evaluating the role of the temporal relationship between health status and leaving work. Counterfactual survival curves among cohorts of exposed (solid) and unexposed (dashed) workers following etiologic (black) and realistic (red) interventions.

Counterfactual survival under no exposure was the same for etiologic and realistic interventions. Differences between etiologic and realistic effects were thus driven by differences between the always exposed interventions. It is not surprising that the two interventions had similar effects when health status had a strong effect on the outcome (scenario 2). The two effects became more alike with increasing effect of intermediate health status on the outcome (results not shown). In this case, workers that experienced the adverse health event were at high risk of experiencing the outcome, whether or not they left work. Leaving work in this setting did not protect workers health. The two effect measures were most dissimilar when health status had no effect on the outcome (scenario 3). Notably, the exposure - outcome relationship was not confounded by health status in this setting, since health status predicted leaving work and future exposure, but not the outcome. Workers that terminated employment due to poor health were therefore not at increased risk for the outcome. The absence of an indirect effect of exposure on the outcome mediated by health status resulted



(a) Exposure predicts health status in all years ($E(t) \rightarrow H(t)$)

(b) Exposure predicts health status only in the first year ($E(1) \rightarrow H(1)$)

Figure 4.7: Evaluating the role of the temporal relationship between exposure and health status. Counterfactual survival curves among cohorts of exposed (solid) and unexposed (dashed) workers following etiologic (black) and realistic (red) interventions.

in small etiologic effects, and even smaller realistic effects.

If workers experience advanced symptoms that lead them to leave work years after an initial diagnosis, as may be the case with the experience of diabetes-related renal disease, etiologic and realistic effect measures will be very similar (scenario 4). In addition to the excess risk for the outcome mediated by health status, the amount of cumulative exposure accrued prior to leaving work would place susceptible workers at high risk for the outcome. However, when susceptible workers terminated employment early for health-related reasons, they were protected from the outcome under study under the realistic intervention. This was reflected by a large etiologic-to-realistic effects ratio in scenario 5. Our scenarios do not represent all plausible observational occupational studies. Instead we aimed to present a selection of extreme cases that illustrate key aspects.

Estimation of causal effects of exposure in the presence of time-varying confounding affected by prior exposure requires the use of g-methods [6, 54, 60, 72], a class of modern statistical estimation approaches that includes inverse probability weighted estimation [73], targeted maximum likelihood estimation (TMLE) [36], g-estimation of structural nested models [74], and g-computation [6]. In addition to being driven by the available data, the choice of intervention whose effects are estimated may depend on the researcher's familiarity or preference for one of these approaches. For example, several recent occupational studies have used the parametric g-formula to evaluate parameters of realistic interventions[56, 57, 67, 75]. G-estimation of structural nested models has been used to evaluate etiologic effects in a number of occupational applications [12, 53, 56, 58, 76–79]. The only published occupational application of the TMLE evaluated etiologic effects [68]; however, the approach is also appropriate for realist effect estimation.

While all aforementioned approaches adjust for time-varying confounding aspect of the HWSE, it is important to consider the implications of parameters estimated by each application. Parameters of interventions that permit leaving work provide estimates of the disease experience in worker populations under hypothetical standards. In contrast, parameters of interventions that require workers to stay employed answer questions about the effects of long-term, sustained exposure, such as may be acquired through a working lifetime. Policy guidelines intended to protect workers health have typically been based on the latter class of effect estimates.

In conclusion, in this first evaluation of the relationship between causal effects of various workplace interventions, we found that differences between the effects are driven by the strength of the mediating health status outcome relationship, timing of the health status employment termination relationship, and how quickly susceptible workers experience symptoms that may lead them to leave the workforce. Care should be taken in distinguishing between effects of real-world interventions in worker populations, and etiologic effects that would have been observed if workers did not self-select out of the workforce. The two classes of interventions have different implications for disease prevention.

	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$	$t = 9$	$t = 10$	$t = 11$	$t = 12$	$t = 13$	$t = 14$	$t = 15$	$t = 16$	$t = 17$	$t = 18$	$t = 19$	$t = 20$
Scenario 1																				
Etiologic, $\psi_e(t)$	0.00	0.00	0.00	0.00	0.01	0.02	0.04	0.07	0.12	0.18	0.26	0.37	0.51	0.69	0.90	1.16	1.46	1.81	2.19	2.61
Realistic, $\psi_r(t)$	0.00	0.00	0.00	0.00	0.01	0.02	0.03	0.05	0.07	0.10	0.13	0.17	0.21	0.26	0.31	0.36	0.43	0.49	0.56	0.63
$R_1(t) = \psi_e(t)/\psi_r(t)$	–	–	–	–	1.00	1.00	1.48	1.59	1.72	1.89	2.06	2.25	2.46	2.69	2.93	3.18	3.44	3.69	3.92	4.13
Scenario 2																				
Etiologic, $\psi_e(t)$	0.00	0.00	0.01	0.02	0.04	0.08	0.16	0.26	0.41	0.61	0.85	1.15	1.48	1.85	2.25	2.65	3.07	3.48	3.89	4.29
Realistic, $\psi_r(t)$	0.00	0.00	0.01	0.02	0.03	0.07	0.11	0.17	0.25	0.35	0.46	0.59	0.74	0.89	1.05	1.22	1.40	1.59	1.78	1.98
$R_2(t) = \psi_e(t)/\psi_r(t)$	–	–	1.00	1.050	1.17	1.30	1.41	1.51	1.62	1.72	1.83	1.93	2.02	2.09	2.14	2.17	2.19	2.19	2.18	2.16
Scenario 3																				
Etiologic, $\psi_e(t)$	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.03	0.04	0.06	0.09	0.13	0.18	0.24	0.33	0.43	0.56	0.72	0.91
Realistic, $\psi_r(t)$	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.12	0.13
$R_3(t) = \psi_e(t)/\psi_r(t)$	–	–	–	–	–	–	1.00	1.47	1.71	1.97	2.21	2.49	2.82	3.20	3.64	4.16	4.74	5.39	6.10	6.86
Scenario 4																				
Etiologic, $\psi_e(t)$	0.00	0.00	0.00	0.00	0.01	0.02	0.04	0.07	0.11	0.17	0.25	0.35	0.48	0.65	0.86	1.12	1.41	1.75	2.13	2.53
Realistic, $\psi_r(t)$	0.00	0.00	0.00	0.00	0.01	0.02	0.04	0.07	0.11	0.18	0.26	0.37	0.50	0.67	0.86	1.08	1.33	1.61	1.90	2.21
$R_4(t) = \psi_e(t)/\psi_r(t)$	–	–	–	–	1.08	0.99	0.96	0.95	0.95	0.95	0.96	0.96	0.97	0.98	1.01	1.03	1.06	1.09	1.12	1.15
Scenario 5																				
Etiologic, $\psi_e(t)$	0.00	0.00	0.01	0.02	0.04	0.06	0.09	0.14	0.20	0.27	0.37	0.49	0.64	0.83	1.05	1.32	1.62	1.97	2.36	2.79
Realistic, $\psi_r(t)$	0.00	0.00	0.01	0.01	0.02	0.03	0.05	0.06	0.08	0.10	0.12	0.15	0.17	0.20	0.23	0.26	0.30	0.34	0.38	0.43
$R_5(t) = \psi_e(t)/\psi_r(t)$	–	–	1.00	1.41	1.60	1.77	1.96	2.18	2.43	2.71	3.03	3.38	3.76	4.17	4.60	5.02	5.44	5.84	6.21	6.53

Table 4.2: Exposure Effect Estimates, and Ratio of Etiologic-to-Realistic Effects, over time.

Chapter 5

Conclusions

This dissertation is concerned with the estimation of causal effects of occupational exposures on incident cancer events in the context of the healthy worker survivor effect. By controlling for time-varying confounding affected by prior exposure, in *Chapter 2* we were able to establish associations between MWF exposure and colon cancer risk that were not achieved using standard analytical techniques in previous reports. Our analysis is the first to support a possible causal relationship between MWFs, particularly straight fluids, and incident colon cancer. Given the ubiquity of occupational exposure to these chemicals, lowering recommended exposure limits may prevent a large number of colon cancers worldwide.

In *Chapter 3* we considered estimation approaches in the presence of left filtering, which occurs when a secondary outcome follow-up is imposed on an existing cohort. We found that approaches that ignore left filtering lead to effect estimates that were biased downward, and the magnitude of the bias increased with increasing incidence of the disease under study, and with increasing proportion of susceptible workers. To estimate survival in the left filtered data we introduced the delayed entry Kaplan-Meier estimator. It combines two known approaches, the adjusted Kaplan-Meier estimator that adjusts for time-varying confounding and informative censoring, and the delayed-entry approaches traditionally used to address left truncation. While neither our method nor any other known analytical methods fully adjust for left filtering, we found that the delayed-entry adjusted Kaplan-Meier resulted in little bias when disease incidence was not too high and the number of latent cancers was small. In addition, the degree of bias was not affected by increases in the proportion of susceptible workers, or the strength of the time-varying confounding. Bias, however, did increase when the disease incidence was doubled.

In a first evaluation of the relationship between causal effects of various workplace interventions, in *Chapter 4* we found that differences between effects of etiologic interventions that prevent leaving work, and the effects of realistic interventions that assign exposure while workers are actively employed, are driven by the strength of the mediating health status -

outcome relationship, timing of the health status - employment termination relationship, and how quickly susceptible workers experience symptoms that may lead them to leave the workforce. Care should be taken in distinguishing between effects of real-world interventions in worker populations, and etiologic effects that would have been observed if workers did not self-select out of the workforce. In most settings, realistic parameters underestimate the disease risk associated with long-term exposure. Policy guidelines intended to protect workers health should be based on etiologic estimates.

We have demonstrated the ability of causal estimators to correct for time-varying confounding of the exposure-outcome relationship in an applied example that suggests a causal relationship between colon cancer and metalworking fluid exposure. The evaluation of continuous exposure response curves through the estimation of parameters of marginal structural models is very deserving of future research. Further, we have proposed an approach that leads to little bias in the presence of left filtering. However, novel estimators that are based only on the observed cases may provide better alternatives and should be investigated. We hope that our discussion of various workplace interventions highlights important considerations for the design and analysis of future studies of occupational hazards.

Bibliography

[1] a J Fox and P F Collier. “Low mortality rates in industrial cohort studies due to selection for work and survival in the industry.” In: *British journal of preventive & social medicine* 30.4 (1976), pp. 225–30. ISSN: 0007-1242.

[2] E S Gilbert. “Some confounding factors in the study of mortality and occupational exposures.” In: *American journal of epidemiology* 116.1 (1982), pp. 177–88. ISSN: 0002-9262.

[3] R R Monson. “Observations on the healthy worker effect”. In: *J.Occup.Med.* 28.6 (1986), pp. 425–433. ISSN: 00961736.

[4] H Michael Arrighi and Irva Hertz-Picciotto. “The Evolving Concept of the Healthy Worker Survivor Effect”. In: *Epidemiology* 5.2 (1994), pp. 189–196. ISSN: 1044-3983. DOI: 10.2307/3702361.

[5] K Steenland et al. “Negative bias in exposure-response trends in occupational studies: modeling the healthy worker survivor effect”. In: *American Journal of Epidemiology* 143.2 (1996), pp. 202–210. ISSN: 0002-9262.

[6] James Robins. “A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect”. In: *Mathematical Modelling* 7.9-12 (1986), pp. 1393–1512. ISSN: 02700255. DOI: 10.1016/0270-0255(86)90088-6.

[7] Maria Gerhardsson De Verdier et al. “Occupational exposures and cancer of the colon and rectum”. In: *American journal of industrial medicine* 22.3 (1992), pp. 291–303.

[8] Yingxu Zhao et al. “Estimated effects of solvents and mineral oils on cancer incidence and mortality in a cohort of aerospace workers”. In: *American journal of industrial medicine* 48.4 (2005), pp. 249–258.

[9] Sally Picciotto et al. “Healthy worker survivor bias: implications of truncating follow-up at employment termination.” In: *Occupational and environmental medicine* 70.10 (2013), pp. 736–42. ISSN: 1470-7926. DOI: 10.1136/oemed-2012-101332.

[10] Jessie P Buckley et al. “Evolving methods for inference in the presence of healthy worker survivor bias.” In: *Epidemiology (Cambridge, Mass.)* 26.2 (2015), pp. 204–12. ISSN: 1531-5487. DOI: 10.1097/EDE.0000000000000217.

[11] Miguel a Hernán, Sonia Hernández-Díaz, and James M Robins. "A structural approach to selection bias." In: *Epidemiology (Cambridge, Mass.)* 15.5 (2004), pp. 615–625. ISSN: 1044-3983. DOI: `10.1097/01.ede.0000135174.63482.43`.

[12] Ashley I Naimi et al. "Estimating the effect of cumulative occupational asbestos exposure on time to lung cancer mortality: using structural nested failure-time models to account for healthy-worker survivor bias." In: *Epidemiology (Cambridge, Mass.)* 25.2 (2014), pp. 246–54. ISSN: 1531-5487. DOI: `10.1097/EDE.0000000000000045`.

[13] Leslie Stayner et al. "Attenuation of exposure-response curves in occupational cohort studies at high exposure levels". In: *Scandinavian Journal of Work, Environment and Health* 29.4 (2003), pp. 317–324. ISSN: 03553140. DOI: `10.5271/sjweh.737`.

[14] Katie M Applebaum, Elizabeth J Malloy, and Ellen a Eisen. "Reducing healthy worker survivor bias by restricting date of hire in a cohort study of Vermont granite workers." In: *Occupational and environmental medicine* 64 (2007), pp. 681–687. ISSN: 1351-0711. DOI: `10.1136/oem.2006.031369`.

[15] Katie M Applebaum, Elizabeth J Malloy, and Ellen A Eisen. "Left truncation, susceptibility, and bias in occupational cohort studies". In: *Epidemiology (Cambridge, Mass.)* 22.4 (2011), p. 599.

[16] Sadie Costello et al. "Metalworking fluids and malignant melanoma in autoworkers". In: *Epidemiology* 22.1 (2011), pp. 90–97.

[17] Ariana Zeka et al. "Risk of upper aerodigestive tract cancers in a case-cohort study of autoworkers exposed to metalworking fluids". In: *Occupational and environmental medicine* 61.5 (2004), pp. 426–431.

[18] Melissa C Friesen et al. "Distinguishing the common components of oil-and water-based metalworking fluids for assessment of cancer incidence risk in autoworkers". In: *American journal of industrial medicine* 54.6 (2011), pp. 450–460.

[19] Deborah Thompson et al. "Occupational exposure to metalworking fluids and risk of breast cancer among female autoworkers". In: *Am J Ind Med* 47.2 (2005), pp. 153–160. ISSN: 02713586. DOI: `10.1002/ajim.20132`.

[20] Ilir Agalliu et al. "Prostate cancer incidence in relation to time windows of exposure to metalworking fluids in the auto industry". In: *Epidemiology* 16.5 (2005), pp. 664–671.

[21] Jun Xie and Chaofeng Liu. "Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data". In: *Statistics in medicine* 24.20 (2005), pp. 3089–3110.

[22] Per Kragh Andersen et al. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

[23] Peter Boyle and J S Langman. "ABC of colorectal cancer: Epidemiology". In: *BMJ: British Medical Journal* 321.7264 (2000), p. 805.

[24] Michael J Glade. "Food, nutrition, physical activity, and the prevention of cancer: a global perspective". In: *American Institute for Cancer Research/World Cancer Research Fund, American Institute for Cancer Research* (1997), pp. 523–526.

[25] Herman F Weindel. "Elements of selecting and using metal-cutting fluids". In: *Tooling and Production* 43 (1981), pp. 66–71.

[26] Kyle Steenland et al. "Dying for work: the magnitude of US mortality from selected causes of death associated with occupation". In: *American journal of industrial medicine* 43.5 (2003), pp. 461–482.

[27] Ellen A Eisen et al. "Mortality studies of machining fluid exposure in the automobile industry I: A standardized mortality ratio analysis". In: *American journal of industrial medicine* 22.6 (1992), pp. 809–824.

[28] Ellen A Eisen et al. "Mortality studies of machining fluid exposure in the automobile industry. III: A case-control study of larynx cancer." In: *American journal of industrial medicine* 26.2 (1994), pp. 185–202. ISSN: 0271-3586. DOI: `10.1002/(SICI)1097-0274(199709)32:3<240::AID-AJIM9>3.0.CO;2-0`.

[29] Melissa C Friesen, Sadie Costello, and Ellen A Eisen. "Quantitative exposure to metalworking fluids and bladder cancer incidence in a cohort of autoworkers". In: *American journal of epidemiology* (2009), kwp073.

[30] Paige E Tolbert et al. "Mortality studies of machining-fluid exposure in the automobile industry: II. Risks associated with specific fluid types". In: *Scandinavian journal of work, environment & health* (1992), pp. 351–360.

[31] José A Bufill. "Colorectal cancer: evidence for distinct genetic categories based on proximal or distal tumor location". In: *Annals of internal medicine* 113.10 (1990), pp. 779–788.

[32] M F Hallock et al. "Estimation of historical exposures to machining fluids in the automotive industry". In: *American journal of industrial medicine* 26.5 (1994), pp. 621–634.

[33] S R Woskie et al. "Size-selective pulmonary dose indices for metal-working fluid aerosols in machining and grinding operations in the automobile manufacturing industry." In: *American Industrial Hygiene Association journal* 55.1 (1994), pp. 20–9. ISSN: 0002-8894. DOI: `10.1080/15428119491019221`.

[34] Ellen A Eisen et al. "Exposure-response models based on extended follow-up of a cohort mortality study in the automobile industry". In: *Scandinavian journal of work, environment & health* (2001), pp. 240–249.

[35] J Schwab et al. "ltmle: Longitudinal targeted maximum likelihood estimation". In: *R package version 0.9* (2014), pp. 1–3.

[36] Mark J van der Laan and Susan Gruber. “Targeted minimum loss based estimation of causal effects of multiple time point interventions”. In: *The international journal of biostatistics* 8.1 (2012).

[37] Mark J van der Laan and Daniel Rubin. “Targeted maximum likelihood learning”. In: *The International Journal of Biostatistics* 2.1 (2006).

[38] Heejung Bang and James M Robins. “Doubly robust estimation in missing data and causal inference models”. In: *Biometrics* 61.4 (2005), pp. 962–973.

[39] Mark J der Laan and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

[40] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. “Estimation of regression coefficients when some regressors are not always observed”. In: *Journal of the American Statistical Association* 89.427 (1994), pp. 846–866.

[41] Mark J der Laan, Eric C Polley, and Alan E Hubbard. “Super learner”. In: *Statistical applications in genetics and molecular biology* 6.1 (2007).

[42] Eric C Polley, Sherri Rose, and Mark J van der Laan. “Super learning”. In: *Targeted Learning*. Springer, 2011, pp. 43–66.

[43] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.

[44] Jerome H Friedman. “Greedy function approximation: a gradient boosting machine”. In: *Annals of statistics* (2001), pp. 1189–1232.

[45] Oleg Sofrygin. *stremr @ github.com*.

[46] R Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.* 2014.

[47] Judea Pearl. “Causal diagrams for empirical research”. In: *Biometrika* 82.4 (1995), pp. 669–688.

[48] Judea Pearl. *Causality: models, reasoning and inference*. Vol. 29. Cambridge Univ Press, 2000.

[49] Donald B Rubin. “Estimating causal effects of treatments in randomized and nonrandomized studies.” In: *Journal of educational Psychology* 66.5 (1974), p. 688.

[50] James M Robins. “Robust estimation in sequentially ignorable missing data and causal inference models”. In: *Proceedings of the American Statistical Association*. Vol. 1999. 2000, pp. 6–10.

[51] Maya Petersen et al. “Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models”. In: *Journal of Causal Inference* 2.2 (2014), pp. 147–185.

[52] Harvey Checkoway et al. "Latency analysis in occupational epidemiology". In: *Archives of Environmental Health* 45.2 (1990), pp. 95–100. ISSN: 00039896. DOI: `10.1080/00039896.1990.9935932`.

[53] Ashley I Naimi, David B Richardson, and Stephen R Cole. "Causal inference in occupational epidemiology: accounting for the healthy worker effect by using structural nested models." In: *American journal of epidemiology* 178.12 (2013), pp. 1681–6. ISSN: 1476-6256. DOI: `10.1093/aje/kwt215`.

[54] James M Robins. "Marginal structural models versus structural nested models as tools for causal inference". In: *Statistical models in epidemiology, the environment, and clinical trials*. Springer, 2000, pp. 95–133.

[55] Maya L Petersen et al. "Diagnosing and responding to violations in the positivity assumption". In: *Statistical Methods in Medical Research* (2010), p. 0962280210386207.

[56] Andreas M Neophytou et al. "Occupational Diesel Exposure, Duration of Employment, and Lung Cancer: An Application of the Parametric G-Formula". In: *Epidemiology (Cambridge, Mass.)* 27.1 (2016), p. 21.

[57] Jessie K Edwards et al. "Occupational radon exposure and lung cancer mortality: estimating intervention effects using the parametric G formula". In: *Epidemiology (Cambridge, Mass.)* 25.6 (2014), p. 829.

[58] Alexander P Keil, David B Richardson, and Melissa A Troester. "Healthy worker survivor bias in the Colorado Plateau uranium miners cohort". In: *American journal of epidemiology* (2015), kwu348.

[59] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2003.

[60] James Robins. "A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods". In: *Journal of chronic diseases* 40 (1987), 139S–161S.

[61] Mei-Cheng Wang, Nicholas P Jewell, and Wei-Yann Tsai. "Asymptotic properties of the product limit estimate under random truncation". In: *The Annals of Statistics* (1986), pp. 1597–1605.

[62] Wei-Yann Tsai, Nicholas P Jewell, and Mei-Cheng Wang. "A note on the product-limit estimator under right censoring and left truncation". In: *Biometrika* 74.4 (1987), pp. 883–886.

[63] Miguel A Hernán. "The hazards of hazard ratios". In: *Epidemiology (Cambridge, Mass.)* 21.1 (2010), p. 13.

[64] Miguel A Hernán, Alvaro Alonso, and Giancarlo Logroscino. "Cigarette smoking and dementia: potential selection bias in the elderly". In: *Epidemiology* 19.3 (2008), pp. 448–450.

[65] Marc G Weisskopf et al. "Biased Exposure–Health Effect Estimates from Selection in Cohort Studies: Are Environmental Studies at Particular Risk?" In: (2015).

[66] Miguel A Hernán et al. "Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology". In: *American journal of epidemiology* 155.2 (2002), pp. 176–184.

[67] Stephen R Cole et al. "Analysis of occupational asbestos exposure and lung cancer mortality using the g formula". In: *American journal of epidemiology* 177.9 (2013), pp. 989–996.

[68] Daniel M Brown et al. "Occupational Exposure to PM2. 5 and Incidence of Ischemic Heart Disease". In: *Epidemiology* 26.6 (2015), pp. 806–814.

[69] Kyle Steenland. "One agent, many diseases: Exposure-response data and comparative risks of different outcomes following silica exposure". In: *American journal of industrial medicine* 48.1 (2005), pp. 16–23.

[70] Harvey Checkoway and Alfred Franzblau. "Is silicosis required for silica-associated lung cancer?" In: *American journal of industrial medicine* 37.3 (2000), pp. 252–259.

[71] Oleg Sofrygin, Mark J van der Laan, and Romain Neugebauer. {*simcausal*}*: Simulating Longitudinal Data with Causal Inference Applications.* 2015.

[72] Ashley I Naimi, Stephen R Cole, and Edward H Kennedy. "An Introduction to G Methods". In: *International Journal of Epidemiology* (2016), dyw323.

[73] James M Robins. "Marginal structural models". In: (1997).

[74] James Robins and Anastasios A Tsiatis. "Semiparametric estimation of an accelerated failure time model with time-dependent covariates". In: *Biometrika* 79.2 (1992), pp. 311–319.

[75] Alexander P Keil and David B Richardson. "Reassessing the Link between Airborne Arsenic Exposure among Anaconda Copper Smelter Workers and Multiple Causes of Death Using the Parametric g-Formula". In: *Environ Health Perspect* (2016).

[76] Sally Picciotto et al. "Hypothetical interventions to limit metalworking fluid exposures and their effects on COPD mortality: G-estimation within a public health framework". In: *Epidemiology* 25.3 (2014), pp. 436–43. ISSN: 1531-5487. DOI: `10.1097/EDE.0000000000000082`.

[77] Jonathan Chevrier, Sally Picciotto, and Ellen a. Eisen. "A Comparison of Standard Methods With G-estimation of Accelerated Failure-time Models to Address The Healthy-worker Survivor Effect". In: *Epidemiology* 23.2 (2012), pp. 212–219. ISSN: 1044-3983. DOI: `10.1097/EDE.0b013e318245fc06`.

[78] Sally Picciotto et al. "Structural nested cumulative failure time models to estimate the effects of interventions". In: *Journal of the American Statistical Association* 107.499 (2012), pp. 886–900.

[79] David C Christiani et al. “Cotton dust and endotoxin exposure and long-term decline in lung function: results of a longitudinal study”. In: *American journal of industrial medicine* 35.4 (1999), pp. 321–331.