Title
Estimating Causal Effects of Occupational Exposures on Lung Health in the Presence of Competing Risks

Author
Combs, Mary

Publication Date
2023

Peer reviewed|Thesis/dissertation

Estimating Causal Effects of Occupational Exposures on Lung Health in the Presence of Competing Risks

by

Mary Ava Combs

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark van der Laan, Chair
Professor Ellen Eisen
Professor Maya Petersen

Summer 2023

Abstract

Estimating Causal Effects of Occupational Exposures on Lung Health in the Presence of Competing Risks

by

Mary Ava Combs

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Mark van der Laan, Chair

In many longitudinal studies researchers are interested in examining failure times associated with a particular risk in settings where subjects may fail due to only one of multiple competing risks. Using targeted learning, we present a method of estimating the cumulative incidence of the multinomial outcome constructed of all possible risks which does not require independence assumptions among the outcomes. We show analytically and with simulation that this method is suitable within the causal roadmap to provide causal contrasts with natural interpretations. We compare the minimal assumptions required for statistically and epidemiologically unbiased interpretations using this method to those required for cause-specific estimation. We apply this method to estimate the causal effect of average yearly cumulative exposure to metal-working fluids on mortality from chronic obstructive pulmonary disease and cancer among a cohort of autoworkers hired between 1938-1985.

Contents

List of Figures

List of Tables

Introduction

In estimating the causal effects of occupational exposure to straight metal-working fluids on Chronic Obstructive Pulmonary Disease (COPD) death, death from cancer can be viewed as a competing risk, an outcome after which the outcome of interest can no longer be observed. If one is interested in the etiologic process by which exposure affects risk of COPD death, then the ideal experiment would include the ability to intervene on all other death process that are unrelated to COPD. However, if the progression of cancer is related to the risk of COPD death, then the natural parameter of interest represents risks from all related causes. Existing techniques depend on the assumption of independence among outcomes, and do not estimate the risk from all potential causes of death. Therefore, we develop a technique which estimates the risk of death from all related causes simultaneously.

Assuming conditional independence between outcomes one can treat a competing risk as a censoring event. This “cause-specific" risk can be interpreted as the risk of death from COPD under a specified exposure regime if 1) there is an intervention to stop the cancer death process, and 2) that intervention does not change the risk of death from COPD. This method keeps the person-time of those who die from a competing risk as though they did not die, but weights it according to the risk of COPD death of those with similar baseline and time-varying covariates up to the time of death but who did not die of the competing risk. This assumes that there is no information in the fact that someone dies from a competing risk. This neither inherently over nor under reports risk because both of the following scenarios are plausible, depending on the particular death processes in question. It would be possible that an underlying, unmeasured health status actually puts individuals at higher risk of both death from cancer and death from COPD. In this scenario, the risk of death from COPD is actually higher among those who die from cancer (had they not died from cancer) and assuming independence between these death processes results in under-estimating the effect of exposure on the risk of death from COPD. It is also possible that the disease processes relating exposure to cancer death and to COPD death are distinct and one's susceptibility to one process strongly indicates protection from the other. In this scenario, those who die from cancer would actually be at lower risk of COPD death (had they not died of cancer) than their alive counterparts, and assuming independence between these death processes results in over-estimating the effect of exposure on the risk of death from COPD.

Chapter 1

Historical Context of Competing Risks

1.1 Motivation

In many longitudinal studies the time from study entry to a particular event is of interest, but subjects may fail due to any one of multiple outcomes. For example, in randomized controlled trials testing the efficacy of a novel chemotherapy to treat cancer, patients may develop other potentially life-threatening medical conditions, possibly as a direct result of the cancer therapy itself. If these patients die from conditions other than cancer, it is no longer possible to observe the effect of the chemotherapy treatment on the risk of death from cancer.

As a motivating example, consider the United Autoworker-General Motors (UAW-GM) cohort study of almost $40,000$ autoworkers [19]. These autoworkers were observed over many years in order to study the effects of factory exposure to metal working fluids (MWF) on mortality due to chronic obstructive pulmonary disease (COPD). A worker's vital status is followed until time of death, which may be attributed to one of multiple competing causes, or administrative censoring at end of follow-up. COPD takes a long time to develop, during which time workers may die from other diseases with shorter latent periods, including one of several specific cancers that have been associated with MWF exposure. Researchers may be interested in comparing the cumulative incidence of COPD after a certain number of years of followup under various static, dynamic, or stochastic exposure regimes.

1.2 Review of Literature

1.2.1 History of Probability Theory for Competing Risks

In epidemiology, competing risks are frequently defined as events which block observation of an outcome of interest. Using this definition it becomes hard to differentiate a competing risk from a "censoring event" or, more specifically, from "informative censoring". In this paper

we instead return to first principles to understand intuitively what is meant when referring to "competing risks", "cause-specific", or "subdistribution" estimates.

In his 1957 article, "The Estimation of the Probability of Developing a Disease in the Presence of Competing Risks" published in *American Journal of Public Health and the Nations Health* Cornfield utilized concepts from existing actuarial life table methodologies to describe the problem of competing risks in practicing epidemiology and to propose a novel solution [13]. He broadly defined a *life table* as a means of documenting a cohort and the subsequent "withdrawing" of elements from that cohort. He defined a *cohort* as a collection of elements with one or more common *characteristics* and a *withdrawal* as an element from that cohort who loses one or more of the defining characteristic. For example, a cohort of alive, disease-free animals could be followed until one or both of these characteristics is lost. At any given point in follow-up, all subjects are either still in the cohort, or have withdrawn because they are 1) alive with the disease of interest, 2) dead with the disease of interest at the time of death, or 3) dead without the disease of interest at the time of death. In this context, Cornfield states that these mutually exclusive categories of key characteristics whose loss leads to withdrawal from the cohort compete with one another. We can derive an intuitive description of *competing risks* as mutually exclusive causes of withdrawal from a cohort. Cornfield hypothesized an experiment in which one is able to completely eliminate causes of withdrawal other than the one of interest, but that even in this nearly impossible context, "... the probability of developing the disease of interest is defined only in the presence of these other competing causes."

This fundamental issue has besieged researchers for years.

The introduction of the product-limit estimator by Kaplan and Meier in their 1958 paper, "Nonparametric estimation from incomplete observations" proved to be widely adopted [27]. This estimator is appropriate in precisely the scenario Cornfield hypothesized about in which all causes of leaving a cohort are eliminated except the one of interest. Kaplan and Meier stipulate the assumption necessary for their estimator to correctly estimate the cause-specific survival as independence between the cause of interest and the other competing causes. This type of loss, that which is independent of the cause of interest has come to be known as "censoring". The product- limit estimator non-parametrically estimates the *cause-specific survival function*, that is, the proportion, $S(t)$, of the cohort whose lifetime (or time until the event of interest) exceeds t in the absence of competing risks without placing assumptions on the form of $S(t)$. Their original framing is as follows: For a sample size of N, first order and label all event times (including events of interest and competing risks), $0 \leq t'_1 \leq t'_2 \leq \ldots \leq t'_N$. Then,

$$\hat{S}(t) = \prod_r \frac{N-r}{N-r+1}$$

where r are the times, t', such that $t'_r \leq t$ and t'_r is an event of interest. The relationship between a cause-specific survival function, $S(t)$ and its corresponding cause-specific cumulative incidence, $\Lambda(t)$ can be shown:

$$S(t) = \exp\{-\Lambda(t)\}$$

And so one can easily convert between survival and cumulative incidence estimates when only one outcome is of interest and all other outcomes are considered independent. The corresponding estimate of the cumulative incidence function is called the Nelson-Aalen estimator and its properties were analyzed separately [33, 1].

The application of these estimators which relied only on time and frequency information (that is, the time from study entry to one of the competing risks and the number of such risks at any given time) proved to challenging. As demonstrated in their 1960 paper, "Competing Exponential Risks, with Particular Reference to the Study of Smoking and Lung Cancer" Berkson and Elveback demonstrate that the quantity of interest and its interpretation can be just as important, if not more important than the selected estimation method [8]. Using formulations originally derived by Neyman and discussed in his 1950 book, *First Course in Probability and Statistics*, Berkson, et al. first distinguishes between *net* and *crude* risks as those when considering a particular risk alone or in the presence of others, respectively. They note that although the total crude rate is independent of the ordering of events, the individual crude rates are not [34]. They then introduce *instantaneous risk* as the risk of a net event in the next infinitesimally small time period among those who have survived up to that time. This quantity has come to be called the *intensity* of a particular event, and, in certain cases with restricted histories, a *hazard*, both notated with $\lambda(t)$. Using the joint likelihood of both events, Berkson goes on to define the maximum likelihood estimators (MLE) of the net risks of each event. Berkson then applies these methods to a prospective study from the American Cancer Society [25] of $200,000$ men in order to study the effect of smoking on the probability of death from one of four causes: lung cancer, other cancer, coronary disease, or other diseases. The authors use a bespoke set up, defining outcomes according to their cause; they propose that non-smokers are subject to death due to diseases unrelated to smoking, and that smokers are additionally subject to deaths due to lung cancer brought on by smoking and death due to other diseases brought on by smoking. Berkson first estimates the net risks of death due to lung cancer, other cancer, coronary disease, and other disease among smokers and among non-smokers, and compares three methods to quantify the disparity in net risk for each of the outcomes between the smoking groups: 1) Their MLE and system of equations method, 2) Differencing the two net risks, and 3) Taking the risk ratio. Of note, the MLE estimates are very close to the simple difference methods. Their findings emphasize the importance of considering competing risks even when reporting net risks or even when a researcher has interest in only one outcome. The risk ratio between smokers and non-smokers for lung cancer is 9.7, while that for coronary disease is 1.6, and for other cancers and other diseases are 1.4 and 1.3, respectively. In a study reporting these results, it could be easy to focus only on the effect of smoking on lung cancer, describing the other effects as minimal. However, looking at the absolute risk, it is much more likely for both smokers and non-smokers to die from any of the other causes. In fact, increase in net risk due to smoking for lung cancer is 0.4%, for coronary disease it is 1.5%, and for other cancer and other diseases it is 0.4% and 0.7%, respectively. While there is an excess 97 deaths due to lung cancer associated with smoking, there is an excess of 369 deaths due to coronary disease associated with smoking. Lung cancer has a longer latent period than most coronary

diseases, so from a competing risk perspective, it is likely that, among those who died from coronary disease, had they not, in fact, died from coronary disease, the smokers would have been more likely to die from lung cancer than the non-smokers. It is only by considering the increased risk smoking creates for coronary disease that we can say the increased risk to lung cancer is likely *attenuated* by the presence of competing risks.

In 1972, Cox published "Regression models and life-tables" in the *Journal of the Royal Statistical Society* in which he observed that under a set of assumptions, one can estimate the association of an exposure or other baseline characteristic with any single cause-specific hazard [16]. Applied researchers began using this technique because of its potentially powerful results. Cox was not the first to utilize the assumption of proportional hazards, but his work combined the survival and failure rate life table methods with regression, creating a powerful way to directly interpret parameter estimates in nearly the same way as one could interpret linear regression estimates. The proportional hazard assumption merely states that the hazard of an event at time t is a product of a function of the explanatory variables and an unknown (and possibly arbitrary) function of time.

$$\lambda(t; z) = \lambda_0(t)\exp(z\beta)$$

This unknown function of time is called the "baseline" or "reference" hazard and refers to the hazard of an event at time t for individuals with a specified referent set of explanatory variable values, i.e. for $z = 0$, $\lambda(t; z) = \lambda_0(t)$. Then, Cox defines a partial likelihood for individual i such that

$$\begin{aligned} PL_i(\beta) &= \frac{\lambda(t;\beta)}{\sum_{j:T_j \geq t} \lambda(t;\beta)} \\ &= \frac{\exp(z_i\beta)}{\sum_{j:T_j \geq t} \exp(z_j;\beta)} \end{aligned}$$

and a maximum likelihood estimator can be defined for β. Cox's method expands to include a particular class of "time-dependent" variables, namely, those which can be expressed as a function of t. Published as a commentary to the 1972 article, Breslow adds an alternative way of writing the likelihood, namely, by estimating β and λ_0 simultaneously by considering their joint likelihood. Then, assuming a constant hazard between failure times, the Breslow estimator provides the non-parametric MLE (NPMLE) for the cumulative baseline hazard.

Many others have contributed to our under standing of longitudinal data including Dabrowska, Greenwood, Heckman, Kalbfleisch, Prentice, and Wellner. In 1993 Andersen, Borgan, Gill and Keiding wrote their textbook, *Statistical Models Based on Counting Processes* which outlines the counting process notation Aalen had rigorously defined in 1978 [1] and which he and and Johansen used to define the "Aalen-Johansen" estimator – a "matrix version" of the Kaplan Meier estimator, thus allow estimation to occur as individuals transition along multiple states [4, 2]. In their textbook, originally published in 1993, Andersen, et al. provide various levels of rigor in defining counting processes to directly define the likelihood of a longitudinal data generating process with multiple end points with the aim

at reaching various audiences [3]. In this paper, we will attempt to similarly use the least notation necessary to convey our messages. Counting process notation allows for rigorous defining of data, data-generating processes, quantities, and models. This language integrates well with the causal roadmap in which clarity between what is known, unknown, and assumed is paramount for proper interpretation. The Neyman-Rubin causal model and the insights from Judea Pearl have created a new set of principles to answer data questions based in causality [46, 44, 35]. Robins expanded on this framework to develop G-computation, a non-parametric method of identifying a quantity of interest in a longitudinal setting with time-varying exposure and time-varying confounding affected by prior exposure, and to propose methods of estimation [41, 42]. Applying the G-computation approach of writing the entire likelihood of a longitudinal data generating process using Andersen's multiplicative intensities allows for flexible definition of quantities of interest, direct assessment of identifiability of those quantities, transparent statistical model selection and causal contrasts.

In 1999 Fine & Gray readdressed the issue of competing risks within the context of Cox proportional hazards models. They motivate their work by pointing out two things: first, as Cornfield had pointed out years prior, that assuming independence among competing risks was fundamentally flawed and secondly that the quantity of interest was usually not the hazard but cumulative incidence [20]. Under Cox proportional hazards models, the transformation of multiple cause-specific hazards to a single cumulative risk is not only laborious when one is only interested in a single cause, but the association between the baseline characteristic and a hazard will differ from its association with the cumulative incidence when estimated this way. To resolve these issues, they propose to estimate "subdistribution" cumulative incidences (and hazards) in such a way that does not require estimating hazards for every competing risk and which has an easily interpretable coefficient parameter with respect to a single cause.

The Fine & Gray subditribution estimator proposed is now in wide use even though 1) It continues to only be valid under proportional hazards, 2) although a statistically unbiased estimator, the subdistribution cumulative incidence is an epidemiologically biased estimand. The interpretation of the true Fine & Gray subdistribution cumulative incidence is unclear – a disparity that can be more easily demonstrated when viewing the issue of competing risks through a causal framework.

Cornfield hypothesized an experiment in which one is able to completely eliminate causes of withdrawal other than the one of interest, but that even in this nearly impossible context, "... the probability of developing the disease of interest is defined only in the presence of these other competing causes." It is this perspective that we adopt in this paper. Three approaches to addressing competing risks are reviewed: treating the competing risk as a censoring event, using proportional hazards models (and the subdistribution estimator [20]), and using the causal roadmap to estimate cause-specific parameters. We will compare the analytic and applied properties of each approach, but first provide the historical context for the development of each. Last, we propose wider use of the causal roadmap approach for estimating cause-specific cumulative incidence over the Fine & Gray estimator of the subdistribution cumulative incidence because it does not rely on the proportional hazards

assumption, is able to utilize modern estimation techniques which account for time-varying exposure and confounding, and, most significantly, it provides an estimand that is epidemiologically unbiased.

1.2.2 Examples from Clinical and Occupational Epidemiology

Analyses that account for competing risks in longitudinal data are used frequently in epidemiological data. For example, an analysis of the risk of coronary heart disease (CHD) among women enrolled in the "Rotterdam Study" identified the population as "frail", meaning at high risk for other causes of death [50]. As such, researchers identified death by any other cause as a competing risk and proceeded to compare hazard ratio and cumulative incidence estimates using three analysis methods. The hazard ratios compared groups based on a number of baseline covariates, primarily whether or not they were on systolic blood pressure lowering medication, and the cumulative incidence estimates risk of CHD after 10 years in the study. The three methods they compared all used proportional hazard models and included first, treating the competing risk as a censoring event, second, estimating cause-specific risks, and lastly, estimating so-called "subdistribution" risks based on those defined by Jason Fine and Robert Gray in their oft-cited 1999 paper, *A proportional hazards model for the subdistribution of a competing risk* [20]. They also demonstrated methods of adapting existing calibration and discrimination techniques to the estimation methods they used. They found that, when estimating both cause-specific and subdistribution proportional hazard models, hazard ratio estimates were similar but found the cause-specific 10-year cumulative risk estimates as to be higher than the subdistribution estimates.

An analysis of the effect of injection drug use on 6-year risk of acquired immunodeficiency syndrome (AIDS) after initiation of combination antiretroviral therapy (ART), demonstrated a non-parametric estimation method while considering death prior to diagnosis as a competing risk [12]. Standardizing to the study population using inverse probability weighting of injection drug use and of censoring [11], researchers adjust for competing risks by estimating cause-specific risks simultaneously. After accounting for confounding, dropout, and competing risks, they found a lower risk difference of development of AIDS comparing injection drug users to non-injection drug and a smaller risk ratio, when compared to the similar crude risk difference and risk ratio. Additionally, they identified the further assumptions required in order to infer causality from their results.

In a study to compare risk of death among individuals on a kidney transplant waitlist with and without panel-reactive antibodies (PRA) where receiving a kidney transplant can be considered as a competing risk, researchers estimate hazard ratios of three end points: mortality, transplantation and a "composite" outcome of either transplantation or death [45]. They then used proportional hazard models to individually estimate the hazard ratio for each of the end points, treating the other endpoints as censoring events, and lastly estimate the subdistribution hazard ratio of mortality considering transplantation as a competing risk. Individual hazard ratio estimates indicate increased PRA to be associated with higher risk of mortality and lower cumulative incidence of transplantation. The estimated subdistri-

bution hazard ratio of mortality also indicated a higher risk among those with PRA, and was higher than the individual hazard ratio of mortality previously estimated. In contrast, researchers found a lower risk for the composite outcome for those with higher PRA levels, underlining the importance of considering separate endpoints separately. Researchers reiterate the arguments made by Bryan Lau, Stephen Cole, and Stephen Gange in their 2009 paper, *Competing Risk Regression Models for Epidemiologic Data*, [30] that when estimating cause-specific hazard ratios using proportional hazard models, "...the analysis inaccurately assumes that mortality and transplantation are independent events." Further, researchers (where "csHR" and "CIF" are used to refer to the cause-specific hazard ratio and the cumulative incidence function, respectively) reiterate that "a csHR > 1 does not necessarily imply that the subdistribution CIF of exposed patients is higher than the subdistribution CIF of unexposed patients." and that therefore "cause-specific hazards models [using proportional hazards models] make it difficult to predict a given patient's probability of dying or undergoing transplantation as a function of their PRA."

An occupational epidemiologic study which addressed the issue of competing risks used the National Study of Coal Worker's Pneumoconiosis data to examine the effect of coal dust exposure on Ischemic Heart Disease (IHD) mortality among US coal miners [29]. Researchers followed a cohort of 9,078 underground workers from 31 mines in 10 states, until either death from IHD or death from pneumoconiosis. Investigators used a job exposure matrix [5] to quantify exposure. Analyses were stratified by region type: anthracite or bituminous. In the anthracite region, a non-montonic exposure-outcome relationship is apparent, indicative of informative competing risks. This was not seen in bituminous regions. Researchers used proportional hazard models to estimate subdistribution hazard ratios of IHD mortality by exposure quartile for each of the three bituminous regions and the anthracite region. These analysis found a similar pattern of non-monotonic ratios by exposure quartile in the anthracite region, but monotonic ratios by exposure quartile in the bituminous regions. Elevated hazard ratios had confidence intervals which did not contain the null value for two of the three bituminous region. In the anthracite region, none of the hazard ratio confidence intervals excluded the null. Authors then also estimated cause-specific hazard ratios from proportional hazard models for the three bituminous regions, comparing exposure quartiles and comparing regions. We assume death by pneumoconiosis was treated as a censoring event. Similar to the subdistribution hazard analysis, cause-specific hazard ratios were elevated monotonically with exposure quartiles with confidence intervals excluding the null. The cause-specific hazard ratios by region warranted similar interpretation to the subdistribution hazard ratios. The authors carefully excluded the anthracite region from cause-specific analysis due to their findings which supported that the competing risk of death from pneumoconiosis was informative. However, without corresponding estimates of risk of death from pneumoconiosis, it is hard to interpret the non-monotonic subdistribution hazard ratios presented for the anthracite region. As such, their conclusion, that exposure to coal dust increases risk of IHD among coal miners, is undercut by their own statement that, "Mortality risk varied by coal region, which may be due to regional variations in the composition of the coal mine dust particulate." Although a novel point – to categorically consider

composition as opposed to exclusively considering particulate size – this conclusion could be interpreted that dust exposure in the anthracite region was not found to be associated with increased risk of IHD mortality. Without a simultaneous analysis of the effect of exposure on risk of pneumoconiosis mortality, the researchers are unable to deduce that in anthracite region it is likely that increased dust exposure increased risk to two causes of mortality.

1.3 General Probability Theory with Competing Risks

The most basic notation for longitudinal data evaluating survival time considers the time to failure random variable, T, with *density* f, *cumulative distribution function* $F(t) = \mathbb{P}(T \leq t)$ and *survival function* $S(t) = \mathbb{P}(T \geq t) = 1 - F(t)$. The *hazard function* is the instantaneous probability of failure conditional on survival up to time t and defined as $h(t) = f(t)/S(t)$. The hazard function can be expressed as the limit as $dt \to 0$ of $\mathbb{P}(T \in [t, t+dt)|T \geq t)/dt$ for continuous time and $\mathbb{P}(T = t|T \geq t)$ for discrete time. Integrating the hazard gives the cumulative hazard, $H(t) = \int_0^t h(u)du$ where, for continuous time, $H(t) = -log\{S(t)\} \implies S(t) = exp\{-H(t)\}$.

Within the framework of longitudinal data with censoring and competing risks, we will consider two settings: first, a single time-point exposure at baseline with baseline confounders, and second, a setting with time-varying exposure and time-varying confounders.

In both settings, consider n workers, with baseline covariates, W, observed prior to exposure, which are then followed until discrete time T or C, whichever comes first. Letting T denote time until failure and let $\varepsilon \in \{1, \ldots, J\}$ denote the type of failure. Without loss of generality, let $J = 2$ as in our example where we can let $\varepsilon = 1$ indicates death by COPD, the outcome of interest, and $\varepsilon = 2$ indicates death by any cancer, a competing risk. A censoring event such as end of follow up or death by an unrelated cause precludes the observation of T, so let $\widetilde{T} = \min(T, C)$ be the time until end of observation and let $\Delta = \mathbb{I}(T \leq C)$ indicate if observation ended due to failure. We assume no ties and, in discrete time, the time ordering $N_1 \to N_2$.

To write the data in a longitudinal format, we introduce *counting processes* so that each worker, regardless of entry time, has a well-defined observation at every time $t = 1, \ldots, \tau$ for a given time τ. Let the counting process $N_j(t) = \mathbb{I}(T_j \leq t, \varepsilon = j)$ jump from zero to one when there is a failure of type j, and let $dN_j(t) = N_j(t + dt) - N(t)$ indicate that a failure of type j occurred in $[t, t + dt)$. Likewise, let $C(t) = \mathbb{I}(C \leq t)$ jump to one when there is a censoring event, and $dC(t) = C(t+dt) - C(t)$ indicate a censoring event between $[t, t+dt)$. Let $N(t) = (N_1(t), \ldots, N_J(t))$ be the vector of J elements at time t from each failure counting process, and let $N_\bullet(t) = \sum_{j=1}^{J} N_j(t)$ be the combined counting process which jumps to one when there is any failure event. Note that $N_\bullet$ is a counting process itself.

For a given variable or process let the overbar, e.g. $\overline{C}(t) = (C(0), C(1), \ldots, C(t))$ denote the vector containing that variable at each time up to and including time t. The *history* at time t is the set of all variables observed up to and including t, denoted as $\mathcal{F}_t$ where $\mathcal{F}_{t^-}$ is understood to mean the history just prior, but not including, time t. For example,

$\mathcal{F}_{t_0} = \{W, \overline{N}(t_0), \overline{C}(t_0)\}$ is the set of baseline covariates, and the entire failure and censoring counting processes up to and including time t_0.

A variable at time t is dependent upon its *parents*, a subset of the history at time t. For example, the parents of $C(t)$ could simply be the censoring process itself at the preceding time, $Pa(C(t)) = (C(t-1))$ since knowing whether or not a person was censored at time $t-1$ would tell you the entire distribution of $C(t)$. Note that this example assumes censoring is independent of baseline covariates and the outcome process for the sake of parsimony.

The *likelihood* of a data-generating process is a specified parameterization of the probability distribution of the data. We seek to write the likelihood for both the single-time-point and for the time-varying exposure settings parameterized by the *intensity*, defined below. To do so, we first notice an important aspect of the counting processes in our example. Workers can experience at most one outcome or censoring event, so we define the counting processes as *degenerate* after such an occurrence. For example, if a worker dies of COPD at time t_0, then at time $t_0 - 1$, $\overline{N}_1(t_0 - 1) = \overline{N}_2(t_0 - 1) = \overline{C}(t_0 - 1) = 0$, at time t_0, $N_1(t_0) = 1, N_2(t_0) = C(t_0) = 0$, and $N_1(t)$, $N_2(t)$, and $C(t)$ are considered degenerate after t_0, i.e. for $t > t_0$, $N_1(t) = 1, N_2(t) = C(t) = 0$.

A counting process is comprised of a predictible portion, Λ, called the *compensator*, and a random portion, M, a martingale [4]. The *intensity* of a counting process is a measure of the rate of change of its predictable part, and is defined as the function, λ, such that $\Lambda(t) = \int_0^t \lambda(u)du$. The intensity can also be defined

$$\lambda(t) = \frac{1}{dt}\mathbb{E}\{dN(t)|\mathcal{F}_{t-}\}$$

which is the expected number of jumps at an infinitesimally small interval around time t given history $\mathcal{F}_{t-}$. An individual worker's intensity at time t can be characterized as a product of their individual hazard rate $h_i(t)$ and the indicator that they are at risk for jumping at time t. Define $Y = \mathbb{I}(T \geq t)$, then the intensity for worker i can be written

$$\lambda_i(t) = \frac{1}{dt}\mathbb{E}\{dN_i(t)|\mathcal{F}_{t-}\} = \frac{1}{dt}Y_i(t)h_i(t)dt$$

and if we assume the hazard rate across individuals are independent we can express the population level counting process

$$\lambda(t)dt = \mathbb{E}\left\{\sum_{i=1}^{n} dN_i(t)|\mathcal{F}_{t-}\right\} = \sum_{i=1}^{n} Y_i(t)h_i(t)dt$$

Furthermore, if we assume the hazard rate across individuals is independent and identical, we can express the population level counting process

$$\lambda(t)dt = \sum_{i=1}^{n} Y_i(t)h(t)dt$$

In discrete time we make observations at regular intervals $t_1^*, \ldots, t_\tau^*$ but only at a subset of these times does a failure or censoring event occur. As such, we can refer to that subset, $t_1, \ldots, t_K$ which may not be evenly spaced but nonetheless index the observations which contribute to the likelihood.

1.3.1 Single time point exposure

In a single time point exposure setting each worker is assigned at baseline to a job with either high or low exposure to metal-working fluids denoted by $A \in \{0,1\}$. Then, we can write the full data as $(W, A, C, T, \varepsilon)$. Letting $N_j^*(t) = (T \le t, \varepsilon = j)$ be the counting processes of COPD death ($j = 1$) and death related to any cancer ($j = 2$) and $C^*(t) = \mathbb{I}(C \le t)$ to be the counting processes of censoring due to end of follow up, or death due to unrelated causes, we can also write the full data in using a longitudinal data structure as $(W, A, N_j^*(t), C^*(t))$ for $(t : t_1, \ldots, t_K)$ at which failure or censoring events occur.

The observed data can be expressed as $O = (W, A, \widetilde{T}, \Delta, \Delta\varepsilon)$ for $\widetilde{T} = \min(T, C)$ and $\Delta = \mathbb{I}(T \le C)$. Using longitudinal data structure, the observed data can be expressed using the counting processes $N_j(t) = \mathbb{I}(T \le t, \varepsilon = j, T \le C)$ for $j \in \{1,2\}$, or $N_j(t) = \mathbb{I}(\widetilde{T} \le t, \varepsilon = j, \Delta = 1)$, and $C(t) = \mathbb{I}(\widetilde{T} \le t, \Delta = 0)$. Then, for $J = 2$, the observed data can be expressed as $O = (W, A, N_1(t), N_2(t), C(t))$ for $(t : t_1, \ldots, t_K)$. Note that we assume the time ordering of $N_1(t) \to N_2(t)$ at any time t. The likelihood for O factorized according to this time ordering is

$$\begin{aligned} p(O) = p_W(W) g(A|W) \prod_{k=1}^{K} \big\{ & \lambda_1(t_k)^{\mathbb{I}(N_1(t_k)=1)} (1 - \lambda_1(t_k))^{\mathbb{I}(N_1(t_k)=0)} \\ & \lambda_2(t_k)^{\mathbb{I}(N_2(t_k)=1)} (1 - \lambda_2(t_k))^{\mathbb{I}(N_2(t_k)=0)} \\ & \lambda_C(t_k)^{\mathbb{I}(C(t_k)=1)} (1 - \lambda_C(t_k))^{\mathbb{I}(C(t_k)=0)} \big\} \end{aligned} \tag{1.1}$$

for the conditional intensities

$$\begin{aligned} \lambda_j(t_k) &= \mathbb{P}(N_j(t_k) = 1 | Pa(N_j(t_k))), j = 1,2 \\ \lambda_C(t_k) &= \mathbb{P}(C(t) = 1 | Pa(C(t_k))) \end{aligned} \tag{1.2}$$

where

$$\begin{aligned} Pa(N_1(t_k)) &= (W, A, (N_1, N_2, C)(t_j) : j = 1, \ldots, k-1) \\ Pa(N_2(t_k)) &= (Pa(N_1(t_k), N_1(t_k)) \\ Pa(C(t_k)) &= (Pa(N_2(t_k)), N_2(t_k)) \end{aligned} \tag{1.3}$$

Many of the conditional intensities in the above likelihood will be degenerate after some time along $t_1, \ldots, t_k$. Specifically, if $Pa(N_1(t))$ includes observing C has happened, or T has happened, then the remaining terms in the likelihood product are all 1. The data is degenerate from then on, as the counting processes do not change value after an event.

We can express the outcome intensities, $\lambda_j(t)$, in terms of *cause-specific hazards*, $\alpha_j(t)$. Recall $Pa(N_1(t_k)) = \{W, A, (N_1, N_2, C)(t_{k-1})\}$ and that $(N_1, N_2, C)(t_{k-1}) = 0 \iff \widetilde{T} \ge t$.

Then,

$$\begin{aligned}
\lambda_1(t) &= \mathbb{E}\{dN_1(t)|Pa(N_1(t))\} \\
&= \mathbb{E}\{dN_1(t)|W, A, \widetilde{T}\} \\
&= \begin{cases} \mathbb{E}\{dN_1(t)|W, A, \widetilde{T} \geq t\} & \text{if } \widetilde{T} \geq t \\ 0 & \text{if } \widetilde{T} < t \end{cases} \\
&= \mathbb{I}(\widetilde{T} \geq t)\mathbb{E}\{dN_1(t)|W, A, \widetilde{T} \geq t\} \\
&= \mathbb{I}(\widetilde{T} \geq t)\alpha_1(t)
\end{aligned}$$

for $\alpha_1(t) = \mathbb{E}\{dN_1(t)|W, A, \widetilde{T} \geq t\}$.

Notice that

$$\begin{aligned}
\alpha_1(t) &= \mathbb{P}(T = t, \varepsilon = 1|\widetilde{T} \geq t, W, A) \\
\alpha_2(t) &= \mathbb{P}(T = t, \varepsilon = 2|\widetilde{T} \geq t, W, A)
\end{aligned}$$

Similarly, $\lambda_2(t) = \mathbb{I}(\widetilde{T} \geq t)\alpha_2$ and $\lambda_C(t) = \mathbb{E}(dC(t)|Pa(C(t))) = \mathbb{I}(\widetilde{T} \geq t)\alpha_C(t)$ for $\alpha_C(t) = \mathbb{P}(C = t|\widetilde{T} \geq t, W, A)$. Note that $\alpha_1(t) = \mathbb{P}(T = t, \varepsilon = 1|T \geq t, W, A)$ if $C \perp\!\!\!\perp (T, \varepsilon)|\{W, A\}$, $\alpha_2(t) = \mathbb{P}(T = t, \varepsilon = 2|T \geq t, W, A)$ if $C \perp\!\!\!\perp (T, \varepsilon)|\{W, A\}$, and $\alpha_C(t) = \mathbb{P}(C = t|C \geq t, W, A)$ if same independent censoring assumption holds.

So the likelihood for a singe observation, O, can be parameterized by $p_W, g_A, \lambda_1, \lambda_2, \lambda_C$ the marginal distribution of W, the conditional distribution of $A|W$ and the intensities for the outcome and censoring counting processes, respectively. Equivalent is the parameterization with respect to the *cause-specific hazards*, α_1, α_2 and the censoring hazard, α_C. We can choose a statistical model such as proportional hazard models to encode what we know about the data- generating distribution, or for convenience.

Note also the intensity $N = N_1 + N_2$, the counting process of T such that $N(t) = \mathbb{I}(T \leq t, \Delta = 1) = N_1 + N_2$ where the intensity, $\lambda(t) = \lambda_1(t) + \lambda_2(t)$ whereas in discrete time, under different time ordering assumptions, $N_1 + N_2$ can jump from 0 to 2.

1.3.2 Time-varying exposure and/or time-varying confounders

In a setting with time-varying exposure and time-varying confounders, an individual worker may change jobs (and therefore exposure) at at given point in time, and their exposure in a given year may be affected by their previous time at work. Let exposure at time t be binary representing high or low exposure and the time-varying confounder, previous time at work, be a continuous percentage. Let $A^*(t)$ and $L^*(t)$ represent the full data for those variables. We can write the full data using a longitudinal data structure as $(W, N_j^*(t), L^*(t), C^*(t), A^*(t))$ for $(t : t_1, \ldots, t_K)$ at which failure or censoring events occur.

The observed data can be expressed using $A(t) = \Delta A^*(t)$ and $L(t) = \Delta L^*(t)$. Then, for $J = 2$, the observed data is $O = (W, N_1(t), N_2(t), L(t), C(t), A(t))$ for $(t : t_1, \ldots, t_K)$.

Note that We assume the time ordering $W \to N(t) \to L(t) \to C(t) \to A(t)$. At times we suppress some of this notation by assuming $W \subset L(0)$, $N(t) \subset L(t)$ and that $C(t) \subset A(t)$. The likelihood for O factorized according to this time ordering is

$$\begin{aligned} p(O) = p_W(W) \prod_{k=1}^{K+1} \big\{ & \lambda_1(t_k)^{\mathbb{I}(N_1(t_k)=1)} (1-\lambda_1(t_k))^{\mathbb{I}(N_1(t_k)=0)} \\ & \lambda_2(t_k)^{\mathbb{I}(N_2(t_k)=1)} (1-\lambda_2(t_k))^{\mathbb{I}(N_2(t_k)=0)} \\ & q_{L(t_k)} \big\} \\ & \prod_{k=1}^{K} \big\{ \lambda_C(t_k)^{\mathbb{I}(C(t_k)=1)} (1-\lambda_C(t_k))^{\mathbb{I}(C(t_k)=0)} \\ & g_{A(t_k)} \big\} \end{aligned} \tag{1.4}$$

for the intensities λ, and conditional densities q and g

$$\begin{aligned} \lambda_j(t_k) &= \mathbb{P}(N_j(t_k) = 1 | Pa(N_j(t_k))), j = 1, 2 \\ q_{L(t_k)} &= \mathbb{P}\{L(t_k) = l(t_k) | Pa(L(t_k))\} \\ \lambda_C(t_k) &= \mathbb{P}(C(t) = 1 | Pa(C(t_k))) \\ g_{A(t_k)} &= \mathbb{P}\{A(t_k) = a(t_k) | Pa(A(t_k))\} \end{aligned} \tag{1.5}$$

where

$$\begin{aligned} Pa(N_1(t_k)) &= (W, (N_1, N_2, L, C, A)(t_j) : j = 1, \ldots, k-1) \\ Pa(N_2(t_k)) &= (Pa(N_1(t_k), N_1(t_k)) \\ Pa(L(t_k)) &= (Pa(N_2(t_k), N_2(t_k)) \\ Pa(C(t_k)) &= (Pa(L(t_k)), L(t_k)) \\ Pa(A(t_k)) &= (Pa(A(t_k)), A(t_k)) \end{aligned} \tag{1.6}$$

Many of the conditional intensities in the above likelihood will be degenerate after some time along $t_1, \ldots, t_k$. Specifically, if $Pa(N_1(t))$ includes observing C has happened, or T has happened, then the remaining terms in the likelihood product are all 1. The data is degenerate from then on, as the counting processes do not change value after an event.

1.3.3 Three Frameworks for Estimation in Settings with Competing Risks

1.3.3.1 Treating a Competing Risk as a Censoring Event

The earliest method to adjust survival and risk estimates in the presence of competing risks has been to consider the competing risk a *censoring event*. There are multiple meanings and ways to interpret the statement that a competing risk is being "treated as a censoring event". In this paper, we refer to a competing event as a censoring event if it is known to

be independent of the outcome of interest, conditional on a set of pre-treatment variables. That is, in the true data generating distribution it is known that subjects who have the censoring event at time t have the same risk of the outcome as those with similar exposure and covariate histories up to time t who were not censored at time t. This condition is frequently called *exchangeability*, *ignorability*, or *coarsening a random*, varying based on the context [23, 26, 21]. In this paper, we say a competing risk is *treated as a censoring event* if the above condition is assumed, but not necessarily known to be true.

Consider data collected between 1962-1977 on 255 patients with malignant melanoma who had a radical operation to remove the tumor and were followed until the end of the study period at which time all subjects had either 1) died from malignant melanoma, 2) were alive on January 1, 1978, the end of the study, or 3) died from other causes [17]. Of the 205 eligible patients, 57 died from the cause of interest, (malignant melanoma), 134 survived until the end of the study, and 14 died from other causes. The outcome of interest, death due to malignant melanoma competes with the remaining two end points. Patients had staggered entry points and the time scale was years since surgery, so being alive at the end of the study period is more of a reflection of when they entered the study rather than their risk of death. It seems reasonable then that risk of death due to malignant melanoma, conditional on baseline characteristics, is similar among those who were alive at the end of the study t years after their surgery as those who were still alive t years after their surgery which was not the end of the study. However, such an argument is less plausible for deaths due to other causes, particularly without further granular data. For example, suppose the incidence of post-surgical infection was related to the aggressiveness of the tumor in a way not captured by baseline characteristics. Then, individuals dying due to surgery-related infections would be at higher risk of death from malignant melanoma at their time of death had they not died from infection, than those who did not die from surgery-related infection at all. In this case, it would be suspect to treat deaths from other causes as censoring. However, if baseline characteristics did fully capture the risk of post-surgical death from malignant melanoma, then, conditional on those characteristics, risk of death from malignant melanoma might be similar among those who did and did not experience death from infection and it might be appropriate to treat that competing risk as a censoring event.

It is a powerful condition, exchangeability, because, if true, treating a competing event as a censoring event allows for identifiability of various quantities of interest including the hazard or cumulative incidence.

Suppose in our motivating example of the autoworker study there are no causes of death other than COPD and cancer, and that all workers are observed until their death. Let T_1 as time until death by COPD, T_2 as time until death by any cancer, and $T = \min(T_1, T_2)$. Let $\Delta = \mathbb{I}(T_2 \geq T_1)$ indicate that T_2 is larger than T_1. Then, the observed data is $O = (T, \Delta)$. Using the typical proof of identifiability with censoring, we can show that under the independence assumption we are able to write the marginal hazard for T_1 at time t, $\lambda^*(t)$, as a function of the observed data.

$$\begin{aligned}
\lambda^*(t) &= \mathbb{P}(T_1 = t | T_1 \geq t) \\
&= \frac{\mathbb{P}(T_1 = t) \cdot \mathbb{P}(T_2 \geq t)}{\mathbb{P}(T_1 \geq t) \cdot \mathbb{P}(T_2 \geq t)} \\
&= \frac{\mathbb{P}(T_1 = t, T_2 \geq t)}{\mathbb{P}(T_1 \geq t, T_2 \geq t)} \\
&= \frac{\mathbb{P}(\min(T_1, T_2) = t, \Delta = 1)}{\mathbb{P}(\min(T_1, T_2) \geq t)} \\
&= \frac{\mathbb{P}(T = t, \Delta = 1)}{\mathbb{P}(T \geq t)} \\
&= \mathbb{P}(T = t, \Delta = 1 | T \geq t) \\
&= \lambda_1(t)
\end{aligned}$$

Provided $T_2 \perp\!\!\!\perp T_1$. Note the last line where the cause-specific hazard is defined.

1.3.3.2 Cox proportional hazards models

In his 1972 seminal paper, Sir David Cox presented the proportional hazards model that specifies the conditional hazard function as the product of an unknown, “baseline" hazard function and an exponential regression function of covariates [16]. He followed up with a detailed discussion of the partial likelihood [15] which, with few alterations, is the Cox proportional hazards model still in use today. An excellent summary of the proportional hazards model and its uses is given by Lin in 2007, and reiterated below [32].

Consider observed data n i.i.d. copies of $O = (\widetilde{T}, \Delta, \{Z(t) : t = 1, \ldots, K\})$ as defined previously, with the addition of the variables $Z(t)$ for $t = 1, \ldots, K$ which are restricted to be dependent upon only t and possibly baseline exposure and covariates. These covariates are often referred to as “time-dependent" covariates but should not be confused with what this paper refers to as “time-varying confounders" [41, 18]. Time-varying confounders refers to covariates dependent upon t, possibly baseline exposure and covariate, and always a non-empty set of exposure and covariates which themselves vary with time and are affected by prior exposure. This distinction is important and will be discussed in more detail later.

The proportional hazards assumption specifies that the hazard function of failure time T, given $Z(t)$, takes the form

$$\lambda(t|Z) = \exp\{\beta Z(t)\}\lambda_0(t)$$

for unknown regression parameters β and arbitrary baseline hazard function λ_0. The partial likelihood proposed to estimate β is then

$$PL(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp\{\beta Z_i(\widetilde{T}_i)\}}{\sum_{j \epsilon \mathcal{R}_i} \exp\{\beta Z_j(\widetilde{T}_i)\}} \right\}^{\Delta_i}$$

for $\mathcal{R}_i = \{j : \widetilde{T}_j \geq \widetilde{T}_i\}$. Taking the derivative we find the score function

$$U(\beta) = \sum_{j=1}^{n} \Delta_i \left\{ Z_i(\widetilde{T}_i) - \frac{\sum_{j \in \mathcal{R}_i} \exp\{\beta Z_j(\widetilde{T}_i)\} Z_j(\widetilde{T}_j)}{\sum_{j \in \mathcal{R}_i} \exp\{\beta Z_j(\widetilde{T}_i)\}} \right\}$$

for which the maximum likelihood estimator for $PL(\beta)$ can be found by solving $U(\beta) = 0$. This gives a parametric estimate, $\hat{\beta}$ for the association between the corresponding baseline covariate (e.g. exposure) and the log hazard of the outcome. Typically, when one is interested not in the hazard but the cumulative incidence of an outcome, one can use the Breslow estimator [16] under the proportional hazards model to simultaneously estimate β and $\Lambda_0(t)$. This method yields the same estimators for β as Cox's method, and the estimator for the baseline cumulative incidence, $\Lambda_0(t)$, can be used to estimate the cumulative hazard (incidence) function at any time. First, the Breslow estimators requires considering $\lambda_0(t)$ as piecewise constant between uncensored failure times. Then, we notice that under the proportional hazards assumption, the cumulative hazard function has a convenient simplification:

$$\begin{aligned} \Lambda(t|Z) &= \int_0^t \lambda_0(u) \exp\{\beta Z\} du \\ &= \exp\{\beta Z\} \Lambda_0(t) \end{aligned}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ is the baseline cumulative hazard function. Then, we write the joint likelihood for β and Λ_0 as

$$L(\beta, \Lambda_0) = \prod_{i=1}^{n} \left\{ \exp\{\beta Z_i(\widetilde{T}_i)\} \lambda_0(\widetilde{T}_i) \right\}^{\Delta_i} \exp\left\{ - \int_0^{\widetilde{T}_i} \exp\{\beta Z_i(t)\} \lambda_0(t) dt \right\}$$

which simultaneously yields the non-parametric maximum likelihood estimators (NPMLE) for β as above and for $\Lambda_0(t)$ as

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \frac{\mathbb{I}(\widetilde{T}_i \leq t) \Delta_i}{\sum_{j \in \mathcal{R}_i} \exp\{\hat{\beta} Z_j(\widetilde{T}_i)\}}$$

To expand this notation for multiple risks we define the cause-specific hazard functions [38]

$$\lambda_j(t|Z) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\{t \leq T < t + \Delta t, J = j | T \geq t, Z(t)\}}{t}$$

for failure types $j = 1, \ldots, J$. The cause-specific hazard, $\lambda_j(t|Z)$, is the instantaneous failure rate at time t from cause j given covariates $Z(t)$ and the existence of all other causes of failure observed. One can consider the overall hazard among J distinct failures as the sum of the cause-specific hazards.

$$\lambda(t|Z) = \sum_{j=1}^{J} \lambda_j(t|Z)$$

The likelihood can be considered in terms of the overall survivor function, $F(t_i|Z_i)$ and the cause-specific hazards.

$$\prod_{i=1}^{n}\{[\lambda_{j_i}(t_i|Z_i)]^{\delta_i}F(t_i|Z_i)\} = \left(\prod_{i=1}^{n}[\lambda_{j_i}(t_i|Z_i)]^{\delta_i}\prod_{j=1}^{J}\exp\left\{-\int_0^{t_i}\lambda_j[u|Z]du\right\}\right)$$

This likelihood factors into separate components for each cause-specific hazard. As seen on above in the treating a competing risk as a censoring event section we saw that the cause-specific hazard identifies the marginal hazard only in the event of independence among the outcomes. This is where many claim that the mere writing of a likelihood as a product of cause-specific likelihoods is "treating a competing risk as a censoring event." However, one need not make that additional assumption, and merely interpret the cause-specific hazards as the quantities they are: the instantaneous risk among at time t of failure by cause j.

Assuming the cause-specific hazard functions satisfy the proportional hazards model, i.e. that

$$\lambda_j(t|Z_j) = \lambda_{0j}(t)\exp\{\beta_j Z_j\}$$

for $j = 1, \ldots, J$ causes of failure. The partial likelihood can be written

$$PL(\beta) = \prod_{j=1}^{J}\prod_{i=1}^{n_j}\left[\frac{\exp\{\beta Z_{ji}(\widetilde{T}_{ji})\}}{\sum_{k=1}^{J}\sum_{l=1}^{n_j}\mathbb{I}(\widetilde{T}_{kl}\geq\widetilde{T}_{ji})\exp\{\beta Z_{kl}(\widetilde{T}_{ji})\}}\right]^{\Delta_{ji}}$$

With corresponding MLE $\hat{\beta}$ and Breslow estimator of the baseline cumulative hazard function

$$\hat{\Lambda}_0(t) = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{\mathbb{I}(\widetilde{T}_{ji}\leq t)\Delta_{ji}}{\sum_{k=1}^{J}\sum_{l=1}^{n_j}\mathbb{I}(\widetilde{T}_{kl}\leq\widetilde{T}_{ji})\exp\{\hat{\beta}Z_{kl}(\widetilde{T}_{ji})\}}$$

Note that only the j^{th} component is needed to estimate the β_j coefficients, and that in such an estimation the other cause-specific hazards do not need to adhere to the proportional hazards assumption, nor are other assumptions needed to be put on the relationship between the outcomes other than mutual exclusivity. Breslow is careful to advise that, "The interpretation of such effects is, however, restricted to actual study conditions and there is no implication that the same regression estimates would prevail under a new set of conditions in which, for example, certain causes of failure have been eliminated." This is only if we wish to identify the marginal hazards with the cause-specific hazards.

To apply the Breslow estimator to obtain cause-specific, cumulative incidence estimates, one needs estimates for all β to obtain the estimate for the baseline cumulative incidence function. Believing these estimates to be true requires the proportional hazards assumption to hold for each cause-specific hazard.

Fine and Gray [20] sought to define a single proportional hazards model from which one could estimate cumulative incidence without relying on models and estimates for the multiple causes of failure. Based on earlier work, [22], Fine and Gray redefine failure time as

a "improper random variable", T^*, as a combination of the constituent failure times. Recall T_1 and T_2 as failure times for risks 1 and 2, respectively, $T = min(T_1, T_2)$ observed failure time and $\varepsilon \in \{1, 2\}$ denoting failure type.

$$T^* = \mathbb{I}(\varepsilon = 1) \times T + \{1 - \mathbb{I}(\varepsilon = 1)\} \times \infty$$

The implied *subdistribution hazard*, $\lambda_1^*(t|Z)$, is then defined as the instantaneous risk of failure type 1 at time t among those who have not had risk 1 or who have had risk 2.

$$\lambda_1^*(t|Z) = \lim_{\Delta \to 0} \frac{1}{\Delta t} \mathbb{P}\{t \leq T \leq t + \Delta t, \varepsilon = 1 | T \geq t \cup (T \leq t \cap \varepsilon \neq 1), Z\}$$

Assuming proportional hazards, $\lambda_1^*(t|Z) = \lambda_{10}^*(t)\exp\{Z(t)\beta_0\}$ the partial likelihood and corresponding MLE for β^*, are constructed. Similarly, a version of Breslow's estimator provides a consistent estimate for $\Lambda_{10}^*(t) = \int_0^t \lambda_{10}^*(u)du$ to estimate the cumulative incidence function. The authors continue by defining an inverse probability of censoring weighted estimator [43] based on the previously defined score equation for instances where censoring events occur and are not necessarily observed. This is contrasted to the case where censoring only occurs due to study end, where the censoring time is observed on all individuals regardless of their observed outcome. For censoring incomplete data, Fine and Gray do recommend an adjusted set of estimators to account for this additional source of uncertainty, which simplify to those already proposed when censoring is completely observed. The authors are the first to point out that this hazard and its associated risk set is "unnatural."

1.3.3.3 Causal Roadmap

The Causal Roadmap integrates causal modeling and statistical estimation [36]. It consists of the following seven steps.

1. Specify knowledge about the system to be studied using a causal model
2. Specify observed data and their link to the causal model
3. Specify the target causal quantity
4. Assess Identifiability
5. Commit to a statistical model and estimand
6. Estimate
7. Interpret

1.4 Causal Roadmap Applied to Competing Risks

1.4.1 Specify knowledge about the system to be studied using a causal model

Consider the two experiments described previously, with n workers with a baseline covariate, W observed prior to exposure, which are then followed until discrete time T, the first occurrence of one of J failures denoted by $\varepsilon \in \{1, \ldots, J\}$. Without loss of generality, let $J = 2$. So, letting T_1 denote time until failure $\varepsilon = 1$, i.e. death by COPD, and T_2 denote time until failure $\varepsilon = 2$, i.e. death by any cancer, define $T = \min_j(T_j)$. Let C denote the discrete time until a censoring event such as death by any other cause or administrative censoring, and let $\Delta = \mathbb{I}(C \geq T)$ and $\widetilde{T} = \min(T, C)$. Since each worker can experience at most one outcome (or neither), we can parameterize each outcome as an *underlying counting process* such that $N_j(t)$ counts the number of failures of type j in $[0, t + dt)$, and $dN_j(t)$ indicates failures of type j in $[t, t + dt)$. $C(t)$ counts the number of censoring events in $[0, t + dt)$. We assume no ties, and, without loss of generality, the time ordering $N_1 \to N_2$.

For both settings we can write structural equations to describe the true underlying distributions and impose restrictions on the model space (the set of all possible distributions of the true data) that are based on domain knowledge. For example, let $\{X, U\} \sim P \in \mathcal{M}^F$ where U is the set of exogenous variables and $X = (Pa(X), U)$ is the set of endogenous variables generated by a set of functions, f, and parents, $\{Pa(X)\} \subset \{X, U\}$. Then $\mathcal{M}^F$ is the structural causal model containing all possible distributions, P, of $\{X, U\}$. We can choose to assume that the error distributions are independent.

In the simple point-exposure setting each worker is assigned at baseline to a job with either high or low exposure to metal-working fluids denoted by $A \in \{0, 1\}$ and the counting processes $N_1(t)$ and $N_2(t)$ and censoring counting process $C(t)$ are followed up for times $t = 1, \ldots, \tau_i$ where τ_i would be the end of followup for individual i (which varies by individual due to staggered entry times). We assume time ordering $C \to N$ at each time.

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ C(t) &= f_{C_t}(W, A, C(t-1)) \\ N_j(t) &= f_{N_j}(W, C(t), U_N) \end{aligned}$$

The information in the above structural equations can be expressed in a directed acyclic graph (DAG) and accompanying list of assumptions – typically restraining the set of unobserved variables.

In a setting with time-varying exposure and time-varying confounders, an individual worker may change jobs (and therefore exposure) at at given point in time, and their exposure in a given year may be affected by their previous time at work. The time-varying exposure is denoted by $A(t) \in \{0, 1\}$ for time t with $\overline{A}(t) = (A(0), A(1), \ldots, A(t))$ denoting the entire history of exposure up to and including time t, and time-varying confounder denoted as $L(t)$

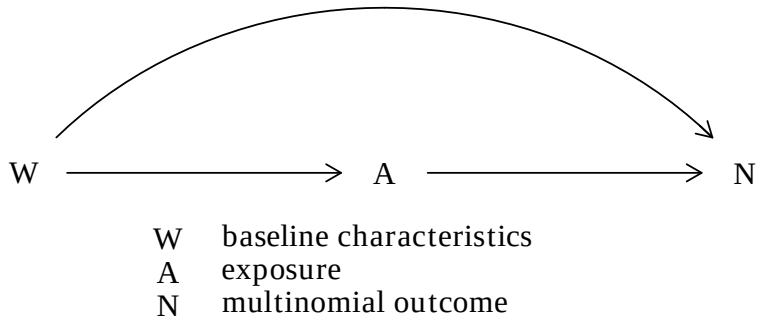

Figure 1.1: Directed acyclic graph for point exposure

for time t with $\overline{L}(t) = (L(0), L(1), \ldots, L(t))$ denoting the entire history of time at work up to and including time t. Define variable after death/censoring equal to last observed value.

We assume a time ordering of $W \to C(t) \to N(t) \to L(t) \to A(t)$ for $t = 1, \ldots, \tau_i$ and that $C(0) = 0$, $L(0) \in W$, $N_j(0) = 0$ for all j.

$$\begin{aligned}
W &= f_W(U_W) \\
A(0) &= f_{A_0}(W, U_A) \\
C(t) &= f_{C_t}(W, \overline{A}(t-1), \overline{L}(t-1), C(t-1), U_C) \\
N_j(t) &= f_{N_{jt}}(W, \overline{A}(t-1), \overline{L}(t-1), C(t), U_N) \\
L(t) &= f_{L_t}(W, \overline{A}(t-1), \overline{L}(t-1), C(t), U_C) \\
A(t) &= f_{A_t}(W, \overline{A}(t-1), \overline{L}(t), C(t), U_A)
\end{aligned}$$

The DAG and corresponding assumptions for the above structural equation model can be seen in figure 1.2.

1.4.2 Specify observed data and their link to the causal model

In the point exposure scenario, the observed data is n i.i.d copies of $O = (W, A, \widetilde{T}, \Delta, \varepsilon) \sim P_0$, or, using counting process notation, the observed data through time K can be expressed as n i.i.d copies of $O = (W, A, \{C(t), dN_1(t), dN_2(t) : t = 1, \ldots, K\}) \sim P_0 \in \mathcal{M}^F$, the set of all possible distributions of P_0.

For the case with time-varying exposure and time-varying covariates, the observed data is n i.i.d copies of $O = (W, \widetilde{T}, \Delta, \varepsilon, \{A(t), L(t) : t = 1, \ldots, K\}) \sim P_0$, or, using counting process notation, the observed data can then be expressed as n i.i.d copies of $O = (W, \{dN_1(t), dN_2(t), C(t), A(t), L(t) : t = 1, \ldots, K\}) \sim P_0 \in \mathcal{M}^F$, the set of all possible distributions of P_0.

If the structural equations defining the causal model accurately describe the data-generating distribution that generated O, then the data are n i.i.d. draws of O from the corresponding system of equations.

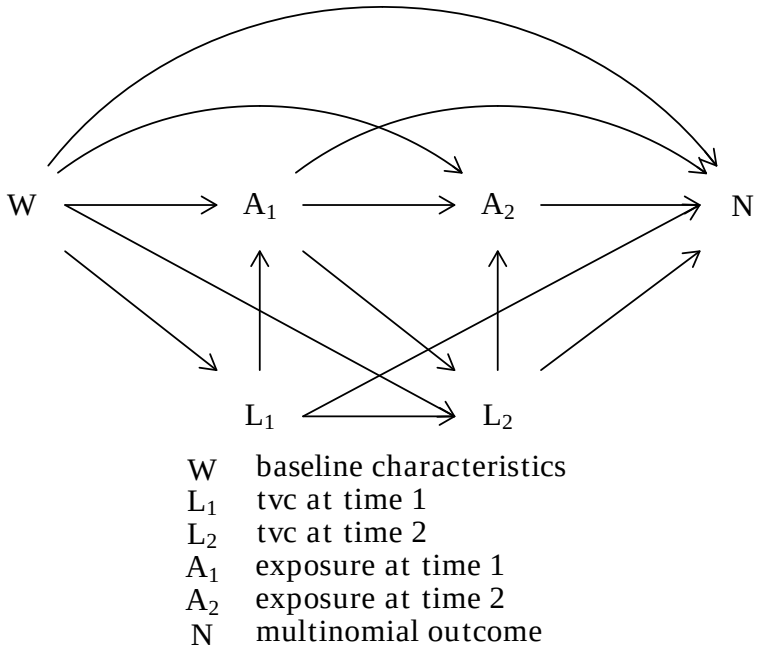

Figure 1.2: Directed acyclic graph for time varying exposure and covariates

1.4.3 Specify the target causal quantity

We will use the formal language of counterfactuals to explicitly state the hypothetical experiment that would generate the full data, including the unobserved counterfactual distribution, and the corresponding target counterfactual quantity as a parameter of the full distribution that would answer the scientific question of interest.

Since the selection of j and K are arbitrary, results apply to estimation of cumulative incidence under other values of $j = 1, \ldots, J$ and $K = 1, \ldots, \max(\tau_i : i = 1, \ldots, n)$.

Under a point treatment setting consider the fixed treatment regime $A = a_0 \in \{0, 1\}$. Suppose we are interested in the cumulative incidence of a particular outcome after some number of years if all workers had been initially assigned a low exposure job compared to that if all workers had been initially assigned to a high exposure job. Recall that, in this scenario, we causally believe that only the exposure at the initial time point is relevant to the structural model. In truth, this is unlikely in an occupational setting, and this example is more fitting to a time-varying data structure. Perhaps a more appropriate example would be in a randomized controlled trial with a single drug administration, or a cohort study of mortality after a specified surgery intervention.

We let N_j^a be the (possibly contrary to fact) counterfactual value of N_j for an individual under treatment regime a. In this case, the target causal quantity is the cumulative incidence

of events of type $\varepsilon = j$ at a fixed time K under a specified treatment, a_0.

$$\Lambda_j^{a_0}(K) \coloneqq \mathbb{P}(N_j^{a_0}(K) = 1) \tag{1.7}$$

Under a time varying setting, consider the static treatment regime $\overline{A} = \overline{a}_0 \in \{\overline{0}, \overline{1}\}$ where exposure level is set to either always high or always low. This structural model seems more appropriate for this example because it is likely that health effects change as exposure accumulates and vice-versa as workers change jobs based on their physical abilities. Suppose we are interested in the cumulative incidence of a particular outcome after some number of years if all workers had been at a low exposure job compared to that if all workers had been at a high exposure job.

We let $N_j^{\overline{a}}$ be the (possibly contrary to fact) counterfactual value of N_j for an individual under treatment regime $\overline{a}$. In this case, the target causal quantity is the cumulative incidence of events of type $\varepsilon = j$ at a fixed time K under a specified treatment, $\overline{a}_0$.

$$\Lambda_j^{\overline{a_0}}(K) \coloneqq \mathbb{P}(N_j^{\overline{a}_0}(K) = 1) \tag{1.8}$$

1.4.4 Assess identifiability

The target causal quantity is identified if it is equal to a parameter of the distribution of the observed data. The parameter of the observed data distribution is called the estimand. In the point treatment setting we can define the cause-specific hazard of event j at time K under treatment regime a_0 and baseline covariates W as

$$\lambda_j^{a_0}(K; W) \coloneqq \mathbb{P}\{dN_j(K) = 1 | C(K) = 0, N(K-1) = 0, A = a_0, W\}$$

and the combined hazard of any event at time K is

$$\lambda_\bullet^{a_0}(K; W) \coloneqq \sum_{j=1}^{J} \lambda_j^{a_0}(K; W)$$

Integrating over the baseline covariate distribution, Q_W and factoring out the survival through time $K-1$ of all causes as 1– the combined hazard at each time leading up, we can express an estimand of the observed data as

$$\mathbb{E}_W\{\Lambda_j(K; W, A = a_0)\} = \int_w \sum_{t=1}^{K} \left[\lambda_j^{a_0}(t; w) \prod_{s=1}^{t-1} \left\{1 - \lambda_\bullet^{a_0}(s; w)\right\} \right] dQ_W(w)$$

The observed data parameter corresponds to the counterfactual cumulative incidence under conditional independence of treatment assignment, positivity, and coarsening at random.

$$\Lambda_j^{a_0}(K; W) = \mathbb{P}\{N_j(K) = 1 | dN(K-1) = 0, C(K) = 0, A = a_0, W\} \tag{1.9}$$

In the time-varying setting we can define the cause-specific hazard of event j at time K under treatment regime $\overline{a_0}$, baseline covariates W, and time varying covariates as

$$\lambda_j^{\overline{a_0}}(K;W,L) \coloneqq \mathbb{P}\{dN_j(K)=1|C(K)=0,N(K-1)=0,A=a_0,W\}$$

and the combined hazard of any event at time K is

$$\lambda_\bullet^{a_0}(K;W) \coloneqq \sum_{j=1}^{J}\lambda_j^{a_0}(K;W)$$

Integrating over the baseline covariate distribution, Q_W and factoring out the survival through time $K-1$ of all causes as 1− the combined hazard at each time leading up, we can express the parameter of interest as

$$\begin{aligned}\Lambda_j^{a_0}(K) &\coloneqq \mathbb{E}_W\{\Lambda_j^{a_0}(K;W)\}\\ &= \int_w \sum_{t=1}^{K}\left[\lambda_j^{a_0}(t;w)\prod_{s=1}^{t-1}\left\{1-\lambda_\bullet^{a_0}(s;w)\right\}\right]dQ_W(w)\end{aligned}$$

The observed data parameter corresponds to the counterfactual cumulative incidence under sequential conditional independence of treatment assignment, positivity, and coarsening at random.

$$\Lambda_j^{a_0}(K;W,L) \coloneqq \mathbb{P}\{dN_j(K)=1|dN(K-1)=0,C(K)=0,A=a_0,W,L(K)\} \tag{1.10}$$

1.4.5 Commit to a statistical model and estimand

1.4.5.1 Specify the statistical model

The likelihood of the observed data, $P_0 \in \mathcal{M}$ can be expressed as a product of the marginal distribution of the baseline covariate, the conditional distribution of exposure given the baseline covariate, the conditional probability of $N_1 = 1$ given the past and no ties and the conditional probability of $N_2 = 1$ given the past and no ties. We assume the outcome ordering of $N_1 \to N_2$, and at this point we place no restrictions on the model, $\mathcal{M}$ containing all possible distributions of P_0.

$$L(K) = \prod_{t=1}^{K} Q_{N(t)}(N(t)|W=w,A=a,N(t-1)) \times g_A(A=a|W=w) \times Q_W(W=w)$$

1.4.5.2 Specify additional convenience assumptions and the corresponding augmented statistical model

This model does not require independence of outcomes but causal assumptions are required for causal inference.

1.4.5.3 Specify the estimand

The estimand is a parameter of the distributions contained in the augmented statistical model that, if the convenience assumptions are true, is equal to the target causal quantity. Our estimand, the cumulative incidence, can be expressed in terms of cause-specific hazards or intensities of the different counting processes. These can be expressed using the counterfactual sub-distributions under interventions on exposure and censoring for each outcome possibilty, marginalized over covariates.

We can write the cause-specific hazard for outcome j at time K under counterfactual intervention a as:

$$\lambda_1^a(w) = \mathbb{E}\{dN_j(1)|A = a, C(K-1) = 0, N(K-1) = n(K-1), W = w\}$$

And we can denote the sum of all hazards at time K as

$$\lambda^a(K; w) = \sum_{j=1}^{J} \lambda_j^a(K; w)$$

Then, the cumulative incidence of outcome 1 at time K can be expressed as

$$\Lambda_1^a(K) = \sum_{t=1}^{K} \{\lambda_1^a(t; w) \times (\prod_{s=1}^{t-1} [1 - \lambda^a(s; w)])\}$$

e.g. for $K = 3$

$$\begin{aligned}\Lambda_1^a(3; w) \; &= \lambda_1^a(1; w) \times (1) \\ &+ \lambda_1^a(2; w) \times (1 - \lambda^a(1; w)) \\ &+ \lambda_1^a(3; w) \times ([1 - \lambda^a(1; w)] \times [1 - \lambda^a(2; w)])\end{aligned}$$

To find the outcome of interest, i.e., the counterfactual cumulative incidence of COPD at time K setting exposure $A = a^*$ over the entire population, we marginalize over the baseline covariate distribution:

$$\Lambda_1^{a^*}(K) = \sum_w [\sum_{t=1}^{K} \{\lambda_1^{a^*}(t; w) \times (\prod_{s=1}^{t-1} [1 - \lambda^{a^*}(s; w)])\}] \times \mathbb{P}(W = w)$$

e.g. for $K = 3$

$$\begin{aligned}\Lambda_1^{a^*}(3) \; &= [\lambda_1^{a^*}(1; w_1) \times (1) \\ &+ \lambda_1^{a^*}(2; w_1) \times (1 - \lambda^{a^*}(1; w_1)) \\ &+ \lambda_1^{a^*}(3; w_1) \times ([1 - \lambda^{a^*}(1; w_1)] \times \; [1 - \lambda^{a^*}(2; w_1)])] \times \mathbb{P}(W = w_1) \\ &+ [\lambda_1^{a^*}(1; w_2) \times (1) \\ &+ \lambda_1^{a^*}(2; w_2) \times (1 - \lambda^{a^*}(1; w_2)) \\ &+ \lambda_1^{a^*}(3; w_2) \times ([1 - \lambda^{a^*}(1; w_2)] \times \; [1 - \lambda^{a^*}(2; w_2)])] \times \mathbb{P}(W = w_2)\end{aligned}$$

1.4.6 Estimate

The estimator is selected typically based on statistical properties. Targeted Minimum Loss-Based Estimation (TMLE) allows for the estimation of cumulative incidence of disease in a cohort following a specified intervention or exposure regime. The longitudinal TMLE was first described by van der Laan and Gruber [28] and utilizes statistical notation first proposed by Bang and Robins [7].

The LTMLE estimator estimates a series of nested conditional regressions for counting down from the year of estimation, K, to 1. Each regression is targeted toward the parameter of interest, so that the estimator of the mean outcome under the specified regime is consistent so long as the exposure mechanism models are consistently estimated, even when the outcome regressions are misspecified.[49].

The final result is a marginal estimate of the cumulative incidence of the outcome among the cohort after K year of follow-up had the specified exposure regime been followed by all workers.

Any data-adaptive method can be used to fit each of the nested regressions. Typically we use SuperLearner [47] to fit main-term logistic regressions for each of the exposure, censoring, and outcome models using the previously specified covariates as measured up to the current time period.

In the case of competing risks, we can separately estimate the cumulative incidence for each outcome, and specify the competing risk as a time-varying covariate in the outcome mechanisms. We then propose standardizing these estimates so that their sum equals one by dividing each by their sum. This linear function of three TMLE estimators is itself a targeted minimum loss-based estimator.

Let $\hat{\Psi}_k$ be the ltmle estimator of parameters Ψ_k for $i = 1, 2, 3$ for the three outcomes: death from COPD ($k = 1$), death from any cancer ($k = 2$), and death from neither COPD or cancer ($k = 3$). Note the influence curve for each estimator is IC_k such that $\hat{\Psi}_k - \Psi_k \approx \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_k(O_i)$.

Now, $\sum_{k=1}^{3} \Psi_k = 1$ but $\sum_{k=1}^{3} \hat{\Psi}_k \neq 1$

Let $\overline{\Psi} = \sum_{k=1}^{3} \Psi_k$ and $\hat{\overline{\Psi}} = \sum_{k=1}^{3} \hat{\Psi}_k$. Note $\overline{\Psi} = 1$ but $\hat{\overline{\Psi}} \neq 1$

So,

$$\begin{aligned}&\hat{\Psi}_1 + \hat{\Psi}_2 + \hat{\Psi}_3 - (\Psi_1 + \Psi_2 + \Psi_3)\\ \approx &\frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_1(O_i) + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_2(O_i) + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_3(O_i)\end{aligned}$$

and

$$\begin{aligned}&\hat{\Psi}_1 + \hat{\Psi}_2 + \hat{\Psi}_3 - 1\\ \approx &\frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_1(O_i) + \mathrm{IC}_2(O_i) + \mathrm{IC}_3(O_i)\end{aligned}$$

Recall $\hat{\Psi}_1 - \Psi_1 \approx \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_1(O_i)$ so $\hat{\Psi}_1 \approx \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_1(O_i) + \Psi_1$ and $\hat{\Psi}_1 - \Psi_1 \approx \Psi_1'(\hat{\Psi}_1 - \Psi_1)$

If IC_j is the influence curve of estimator Ψ_{n_j} of parameter Ψ_{0_j}, for each $j = 1, \ldots, d$ then the influence curve of the function $f(\Psi_{n_j} : j$ is an estimator of $f(\Psi_{0_j} : j)$ and $f(\Psi_{n_j} : j) = \sum_j \frac{d}{d\Psi_{0_j}} f(\Psi_0)\text{IC}_j$

1.4.7 Interpretation

Interpretation of results starts by a clear distinction of statistical and causal interpretations. The true value of an estimand may differ from the true value of the causal quantity of interest based on the veracity of the augmented assumptions. The choice of the statistical estimator does not affect the difference between the estimand and the causal quantity of interest, however. Statistical bias is determined by choice in estimator and can be minimized and quantified (within levels of certainty). Sensitivity analyses can attempt to quantify the difference between the estimand and the causal quantity of interest. Plots and tables can be generated to understand the pre-determined research question that can be answered by non-parametric estimates of the cumulative incidence of a particular outcome.

Chapter 2

Applied Estimation of Causal Effects of Occupational Exposure on Competing Causes of Mortality

2.1 Introduction

Estimating the causal effect of high, chronic, occupational exposure to metal-working fluids (MWF) on the risk of death by Chronic Obstructive Pulmonary Disease (COPD) is of interest to occupational health scientists. An analysis based on applying g-estimation in the UAW-GM cohort study suggests lowering the recommended limits of exposure to straight, soluble, and synthetic MWF relative to the 1998 National Institute for Occupational Safety and Health recommendations could save hundreds of years of life lost to COPD, with the greatest saved under a ban on soluble fluids (1550 years) and straight fluids (737 years) [37]. This paper adds to the existing literature by considering competing risks in the same Autoworker study. That study was jointly funded in 1985 by labor and management as a cancer mortality study with an extensive exposure assessment component, motivated by worker concerns about digestive and respiratory cancers in relation to MWF exposure [19].

Soluble fluids are the most common MWF exposure, but straight fluids are classified as human carcinogens by the International Agency for Research on Cancer (IARC) [40]. This paper adds to existing literature by characterizing the cumulative incidence of COPD death due to straight MWF exposure with the additional consideration of the competing risk of cancer death.

COPD takes a long time to develop, during which time individuals under study may die from other diseases with relatively shorter latent periods, such as cancer. It has also been demonstrated in previous literature that MWF exposure has a direct effect increasing risk of death due to cancer in this population [10], largely due to specific cancers including prostate and lung, among others.

The two predominant frameworks for estimating the cumulative incidence of an outcome

in the presence of competing risks are the "cause-specific" and the "subdistribution" [30]. The cause-specific cumulative incidence function describes the marginal risk of the outcome relative not to survival but to the union of survival or the competing risk. When this parameter is estimated using Cox proportional regression models, one typically estimates the cause-specific hazard and transforms this to an estimate of more epidemiologically useful parameter, the cause-specific cumulative incidence [9]. This estimation technique requires both the proportionality assumption of the cause-specific hazard and the assumption that the competing events are independent of each other (so as to "treat the competing risk as a censoring event"). Divorced from the estimation technique, however, the cause-specific cumulative incidence function remains the desired parameter of interest by epidemiologists.

Methods to estimate the subdistribution parameters have been developed to circumvent the independence assumption necessary for interpreting Breslow estimators. The subdistribution hazard and subdistribution cumulative incidence can be unbiasedly estimated, though their interpretation continues to be cause for confusion [6]. Subdistribution parameters are defined via an improper random variable which classifies those who experience the competing risk as having infinite survival time [20]. The convenience of these parameters is that the subdistribution cumulative incidence can be estimated directly from a proportional hazard model.

In this paper we estimate the outcome distribution of the cumulative incidence of death from COPD under high- and low-exposure causal interventions using the non-parametric longitudinal targeted minimum loss-based estimator (ltmle) [31]. We index causal exposure interventions by the observation year in which cumulative exposure switches from low to high as defined by a specified threshold. The thresholds will be pre-selected percentiles from the distribution of cumulative exposure values over the entire cohort and observation time. We anticipate that an earlier switch time increases the risk of COPD, and that interventions that use a higher threshold to define high exposure will result in higher estimated COPD risk.

By directly estimating the cumulative incidence distribution of the outcome we can gain a more complete understanding of impact that an OSHA policy (federal or state) or else informal industry guideline designed to lower MWF exposure in the auto industry would have had on the risk of death due to COPD in this population.

2.2 Methods

2.2.1 Study Population

The United Autoworkers-General Motors (UAW-GM) cohort mortality study was designed to assess the effects of metalworking fluids on workers' health and has been described in detail previously [19]. The study cohort for the current paper was drawn from this larger population of 45,000 automobile production workers from three manufacturing plants in Michigan. The study cohort included 39,309 workers hired between 1938 and 1982 after

restricting eligibility to those with at least three years of employment at GM. Individual follow-up began after three years of work and lasted through 1995.

2.2.2 Exposure

The exposure to metalworking fluids measured in the UAW-GM cohort mortality study has been published previously [24]. Concentration of straight, soluble, and synthetic metalworking-fluid was estimated based on particulate matter (PM) measurements taken from air samples collected between 1958 and 1987 by industrial hygienists. Exposure factors were then assigned to workers based on their plant, department, and specific job obtained from their employment records. Exposure was characterized as one of two a binary variables, where individual workers' time-varying cumulative straight MWF exposure was classified as either above (high) or below (low) the 75^{th} and 90^{th} percentiles of the distribution of cumulative straight MWF exposure in the entire cohort.

Exposure accumulates over time, so a worker with low exposure over a long period of time might have the same cumulative exposure as a worker with high exposure over a short period of time. It would be desirable for analyses to differentiate between workers with the same cumulative exposure but with different exposure histories.

In an effort to make such a distinction, exposure interventions are indexed by the observation year in which the cumulative straight metal working fluid switched from low to high, as defined by the 75^{th}, and the 90^{th} percentiles of all observed cumulative exposure values. Exposures can therefore be described by the "switch time", $S = 0, 1, 2, 3, \cdots, 15$, and by their "threshold" (the 75^{th} or 90^{th} percentiles).

2.2.3 Outcome and Competing Risk Measure

Vital status data were obtained through the Social Security Administration, the National Death Index, plant records, state mortality files, and in some instances death certificates. Cause of death was obtained from ICD-9 and ICD-10 codes from state vital records and death certificates. For this study, deaths were categorized using as either caused by any cancer (ICD-10: C00-D49; ICD-9: 140-165, 170-175, 179-208, 210-239), chronic obstructive pulmonary disease (COPD) (ICD-10: J40-J47; ICD-9: 490-492, 496), or any other cause.

2.2.4 Covariates

Baseline covariates include race (black, white, or unknown), sex (male or female), plant location (plant 1, plant 2, or plant 3), categorical calendar year of hire grouped by quantiles (1938-1952, 1953-1965, 1966-1973, 1974-1981), and binary age of hire (less than 25, or greater than or equal to 25 years old), as well as cumulative exposure to straight metal-working fluid in the first three years at work. This demographic and work history information was obtained through employment records provided by the employer and the labor union.

Time-varying covariates include employment status (actively at work or left work), cumulative time off work (using the 45^{th} percentile as the cut point), average annual exposure to straight MWF (using the 90^{th} percentile as the cut point), and the diagnosis of any cancer up to that year of follow-up. The Michigan Cancer Registry provided ICD-O codes and dates of incident cancer diagnosis for individuals alive during or after 1985, and the use for this autoworkers cohort has been described previously [14].

2.2.5 Censoring

Workers are considered censored if they are still alive in 1995 (administrative censoring), if their observed age exceeds 108, or if they die due to a cause other than cancer or COPD.

2.2.6 Statistical Methods

Targeted Minimum Loss-Based Estimation (TMLE) allows for the estimation of cumulative incidence of disease in a cohort following a specified intervention or exposure regime. We are interested in the effect of high cumulative MWF exposure as defined by the 25th and 75th percentiles on death from COPD after 50 years of followup. Additionally, we compared the effect of delaying the time until a worker switches from low to high cumulative exposure.

We estimated the cumulative incidence of death from COPD in the worker population if, during the first $S-1$ years of follow-up, they all had cumulative MWF exposure below the cut-off, then, during the S through 50^{th} year of follow-up, they all had cumulative MWF exposure above the cut-off. We compared these estimates across multiple values of S and using the two exposure cut-offs. All exposure regimes intervene to prevent censoring.

The longitudinal TMLE was first described by van der Laan and Gruber [28] and utilizes statistical notation first proposed by Bang and Robins [7]. Our target parameter of interest is the mean cumulative incidence of COPD death at each specified time point $t = 1, \ldots, 50$, among workers following the defined exposure regime $\overline{s} = (\overline{0}, \overline{1_{s,\ldots,50}})$ where $t = 1$ indicates the first year of follow-up, and s indicates the follow-up year of switching from low to high cumulative exposure. For instance, following exposure intervention regime $\overline{10} = (\overline{0}, \overline{1_{5,\ldots,50}})$ represents cumulative exposure to MWF below the cut-off for the first 4 years of follow-up, then cumulative exposure to MWF above the cut-off for the fifth through the fiftieth years of follow-up and being uncensored at each time point. The target parameter is $\mathbb{E}_{Y_{\overline{s}}(t)}$ for $a = 0, \ldots, 10$ where $Y(t)$ is an indicator of death from COPD prior to year t and the $\overline{s}$ subscript indicates that this is a counterfactual outcome under intervention regimen $\overline{s}$. Each estimation procedure was performed separately for each intervention regime, exposure cut-off and time $t = 1, \ldots, 50$.

First, for a given t, we estimated the true exposure assignment mechanism, or the probability of workers following the specified intervention regime at time point k, given the observed past, including the treatment and censoring history and other measured covariates, which we denote $g_E(k)$. We similarly generated an estimate for the censoring mechanism $g_C(k)$, which estimated the probability of a worker fitting one of the censoring criteria in

time period k given the observed past. For each of the 22 estimation procedures (eleven switch times and two binary cumulative exposure definitions), fits of the g models g_E and g_C were generated using all person-years where, at time k, the worker had followed the regimen of interest up to time point $k-1$.

We then estimated a series of nested conditional regressions for k counting down from $k = t, \ldots, 1$. Each estimate predicted the probability of the outcome between time points k and t by predicting either the outcome of the previous $(k+1)^{th}$ regression or the observed outcome.

Each regression was targeted toward the parameter of interest, so that the estimator of the mean outcome under the specified regime is consistent so long as the g models are consistently estimated, even when the outcome regressions are misspecified.[49]

The final result is a marginal estimate of the cumulative incidence of the outcome among the cohort after t year of follow-up had regime $\overline{s}$ been followed by all workers.

We used SuperLearner [47] to fit main-term logistic regressions for each of the exposure, censoring, and outcome models using the previously specified covariates as measured up to the current time period. We fit these regressions iteratively for each time period ($t = 1, \ldots, 50$) and generate outcome estimates for the cumulative incidence of death from COPD or cancer among the population at time t. We used these 50 estimates to generate a marginal influence curve to estimate the outcomes distribution over the entire follow-up time. We calculated risk ratios and differences, and their corresponding confidence intervals, using influence curve based variance estimates, and compared across exposure regimes. We used the ltmle package [31] in R version 4.1.2. [39]

2.2.7 Parameter of Interest and Estimators

In a world of no censoring, under each previously specified exposure regime, we wish to estimate the distribution of the cumulative incidence of death from COPD. We use the longitudinal targeted minimum loss-based estimator (ltmle) to estimate separately the cumulative incidence of death from COPD, death from any cancer, and death from either COPD or cancer. We make these estimates at 50 and 55 years of follow-up (the 85^{th} and 90^{th} percentiles of distribution of length of follow-up), and estimate each of them under each exposure regime.

We consider exposure interventions indexed by the observation year in which the cumulative straight metal working fluid switching from low to high, and by the threshold used to define "high" cumulative exposure. Exposures can therefore be described by the "switch time", $S = 0, 1, 2, 3, \cdots, 11$, and by their "threshold", $\theta = 25, 50, 75, 90$.

These estimates will be standardized so that their sum equals one by dividing each by their sum. This linear function of three tmle estimators is itself a targeted minimum loss-based estimator.

Let $\hat{\Psi}_k$ be the ltmle estimator of parameters Ψ_k for $i = 1, 2, 3$ for the three outcomes: death from COPD ($k = 1$), death from any cancer ($k = 2$), and death from neither COPD or cancer ($k = 3$). Note the influence curve for each estimator is IC_k such that $\hat{\Psi}_k - \Psi_k \approx \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_k(O_i)$.

Now, $\sum_{k=1}^{3} \Psi_k = 1$ but $\sum_{k=1}^{3} \hat{\Psi}_k \neq 1$

Let $\overline{\Psi} = \sum_{k=1}^{3} \Psi_k$ and $\hat{\overline{\Psi}} = \sum_{k=1}^{3} \hat{\Psi}_k$. Note $\overline{\Psi} = 1$ but $\hat{\overline{\Psi}} \neq 1$

So,

$$\begin{aligned}&\hat{\Psi}_1 + \hat{\Psi}_2 + \hat{\Psi}_3 - (\Psi_1 + \Psi_2 + \Psi_3)\\ \approx &\frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_1(O_i) + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_2(O_i) + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_3(O_i)\end{aligned}$$

and

$$\begin{aligned}&\hat{\Psi}_1 + \hat{\Psi}_2 + \hat{\Psi}_3 - 1\\ \approx &\frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_1(O_i) + \mathrm{IC}_2(O_i) + \mathrm{IC}_3(O_i)\end{aligned}$$

Recall $\hat{\Psi}_1 - \Psi_1 \approx \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_1(O_i)$ so $\hat{\Psi}_1 \approx \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_1(O_i) + \Psi_1$ and $\hat{\Psi}_1 - \Psi_1 \approx \Psi_1'(\hat{\Psi}_1 - \Psi_1)$

If IC_j is the influence curve of estimator Ψ_{n_j} of parameter Ψ_{0_j}, for each $j = 1, \ldots, d$ then the influence curve of the function $f(\Psi_{n_j} : j$ is an estimator of $f(\Psi_{0_j} : j)$ and $f(\Psi_{n_j} : j) = \sum_j \frac{d}{d\Psi_{0_j}} f(\Psi_0)\mathrm{IC}_j$

2.3 Results

We used a cohort of $n = 39,309$ autoworkers from the United Auto Workers–General Motors (UAW-GM) dataset [19]. We included person-years for each individual starting with their year of hire and ending with their year of death or 2015, whichever came first. Follow-up cancer diagnoses observation started in 1985 and ended in 2016. Follow-up for mortality started in 1941 and ended in 2015. Exposure data ended in 1994. We excluded individuals missing more than half of their work history. We excluded 8 workers with age less than 16, and 4 workers with no observed reason for end of observation.

Characteristics of the cohort used are presented in table 2.1. There were 947 deaths due to COPD, 5,610 and deaths due to any cancer, and the remaining 32,752 were either lost to follow-up or still alive by the end of mortality follow-up in 2015. Follow-up observation lasted for an average of 39 years per worker, and individual workers stayed at work for an average of 17.6 years. The cohort workers contained 34,515(87.8%) male, and 23,295(59.3%) white. There were 19,136(48.7%) workers hired 25 years old or higher, and 18,535(47.2%) workers hired in 1965 or later. On average, workers took 2.6% time off annually, and 6,941(17.7%) were diagnosed with cancer between 1985 – 2016. Across all person-years, the mean average annual straight metal-working fluid exposure was 0.050 mg/m^3, and with median and inter-quartile range (IQR) both 0. Excluding values of 0, the mean average annual straight MWF exposure was 0.473 mg/m^3, with median 0.104 mg/m^3 with IQR 0.346.

The distribution of cumulative straight metal-working fluid exposure across all observed person-years is presented in table 2.2 for the full cohort, and for the subcohorts of those who

	Study population (n=39,309)			
Characteristic	**No.**	**%**	**Mean**	**Median (IQR)**
Follow-up years				39 (13)
Years at work			17.6	
Avg annual straight MWF exp. (mg/m^3)				
all			0.050	0 (0)
non-zero			0.473	0.104 (0.346)
Demographics				
male	34,515	87.8		
white	23,295	59.3		
>= 25 yo at hire	19,136	48.7		
>= 1965 yr of hire	18,535	47.2		
Time-varying covariates				
average time off		2.6		
cancer Dx	6,941	17.7		
Outcomes of Interest				
censored	32,752	83.3		
cancer death	5,610	14.3		
COPD dth	947	2.4		

Table 2.1: Demographic, time-varying covariates, and outcomes of the United Autoworkers-General Motors (UAW-GM) cohort.

die from COPD and those who died from any cancer. In the full cohort, the mean exposure was 2.26 mg/m^3, with median 0.056 mg/m^3 and IQR 0.85. Excluding 0 values, the mean exposure was 4.09 mg/m^3, with median 0.69 and IQR 2.29. In the COPD mortality cohort, the mean exposure was 3.07 mg/m^3, with median 0.09 and IQR 1.40. Excluding 0 values, the mean exposure was 5.45 mg/m^3, with median 1.03 and IQR 3.64. In the cancer mortality cohort, the mean exposure was 2.74 mg/m^3, with median 0.05 and IQR 1.06. Excluding 0 values, the mean exposure was 4.94 mg/m^3, with median 0.86 and IQR 2.92. The 75^{th} and 90^{th} percentiles of the cumulative straight MWF exposure distribution for the full cohort are 0.85 and 4.21 mg/m^3, respectively.

The number of deaths from COPD and cancer by year of follow-up is presented in table 2.3 and in figure 2.1. Most (87.3%) COPD deaths occurred within 55 years of follow-up, and most (94.9%) cancer deaths occurred in the same follow-up time. Table 2.3 and figure 2.2 show the percent of deaths from COPD and from cancer among workers eligible (alive and still being followed) by follow-up year.

The number of COPD and cancer deaths increase between follow-up years $0-40$ and then decrease through follow-up year 75. However, the percent of deaths per observation year among those who were eligible increases throughout the years of follow-up, only decreasing

	Mean	Median (IQR)	75^{th}	90^{th}
Full cohort (n=5,610)				
all exp.	2.26	0.05 (0.85)	0.85	4.21
non-zero	5.45	1.03 (3.64)	3.94	12.9
COPD (n=947)				
all exp.	3.07	0.09 (1.40)	1.4	7.12
non-zero	5.45	1.03 (3.64)	3.94	12.9
Cancer death (n=5,610)				
all exp.	2.74	0.05 (1.06)	1.05	5.32
non-zero	4.94	0.86 (2.92)	3.17	11.04

Table 2.2: Cumulative straight metal-working fluid exposure distribution across UAW-GM cohort, and among those who died from COPD and from cancer.

again after 55 years of follow-up.

The observed number and percent of deaths from COPD and from cancer by switch time exposure regime is presented in table 2.4. The percent represents the proportion of the total number of workers whose exposure history followed a particular exposure regime and who were eligible (still alive and under observation) for the outcomes. Recall that exposure regimes are defined by the switch time, S, at which a worker's cumulative straight metal-working fluid crosses a specified threshold, selected as the 75^{th} and 90^{th} percentiles of the cumulative exposure distributions. We examine 16 switch time regimes per high exposure threshold and compare them to a "never switch" exposure regime in which a worker's cumulative exposure never crosses the 25^{th} percentile threshold. For example, table 2.4 shows that among those whose cumulative straight MWF exposure crossed the 75^{th} percentile threshold, 0.85 mg/m^3, during their first year of observation, S = 1, there were 17 deaths from COPD, and 74 deaths from cancer, which make up 3.9% and 16.9% of the 438 workers eligible and who followed the S = 1 exposure regime. For comparison, among those who never crossed the 75^{th} percentile threshold, 185, (54%) and 4,657, (13.6%) died from COPD and cancer, respectively.

Table 2.4 shows that under each high exposure threshold, among those who died from COPD and from cancer and whose cumulative straight MWF exposure crossed the respective threshold within the first 15 years of observation (the columns marked "Total eligible" in table 2.4), most, 63.6% from COPD and 52.2% from cancer for the 75^{th}, and 50.6% from COPD and 51.8% from cancer for the 90^{th}, are accounted for by those whose cumulative straight MWF exposure exceeded the respective threshold during the "baseline" first three years of their employment (switch time S = 0) or within the first three years of follow-up (switch time S = $1, \ldots, 3$). This trend is lessened, but still present even when we look at the percent of eligible workers by threshold and switch time.

Table 2.5 and figures 2.3 and 2.4 show the TMLE estimated cumulative incidence and 95%

Cause of death	**COPD**	**Cancer**	
Follow-up year	**N (%)**	**N (%)**	**Total eligible**
0-5	2 (0.01)	57 (0.15)	39,309
5-10	3 (0.01)	150 (0.39)	38,948
10-15	15 (0.04)	224 (0.59)	38,264
15-20	41 (0.11)	311 (0.83)	37,294
20-25	32 (0.09)	474 (1.32)	36,003
25-30	93 (0.27)	677 (1.98)	34,197
30-35	111 (0.35)	825 (2.59)	31,857
35-40	152 (0.56)	831 (3.04)	27,327
40-45	147 (0.82)	746 (4.16)	17,932
45-50	145 (1.19)	593 (4.87)	12,168
50-55	86 (1.38)	383 (6.14)	6,241
55-60	81 (2.2)	245 (6.66)	3,677
60-65	28 (1.69)	78 (4.71)	1,657
65-70	9 (2.19)	14 (3.41)	411
70-75	2 (2.56)	2 (2.56)	78

Table 2.3: Observed number and percent of deaths from COPD and from cancer in the UAW-GM cohort among those still eligible by year of follow-up, .

confidence intervals of COPD and cancer deaths after 55 years of follow-up by switch time regime, $S = 0, 1 \ldots, 15$, and by high cumulative straight MWF exposure threshold of 75^{th} and 90^{th} percentiles. These estimates represent the cumulative incidence of COPD death (and death from cancer) if the exposure history for every worker was intervened on and set to the exposure regime of interest. For example, we estimate that if every worker had cumulative straight MWF exposure cross the 75^{th} percentile exactly in their fifth year of observation, $7.8\%(7.4, 8.1)$ would die from COPD, and $21.8(20.5, 22, 7)$ would die from cancer, by the 55^{th} year of follow-up. The "never switch" row shows the estimated cumulative incidence and 95% confidence intervals of COPD and cancer deaths after 55 years of follow-up if all workers had cumulative straight MWF exposure levels below the 25^{th} percentile for all follow-up.

A visualization of the TMLE estimates for the cumulative incidence of COPD and cancer death after 55 years of follow-up by switch time regime and by high straight MWF exposure threshold is presented in figures 2.3 and 2.4.

In figure 2.3 we see an outlier at switch time $S = 11$ and high exposure threshold 90^{th}, and an overall increase in variability in estimates for switch time regimes for $S \geq 10$. Otherwise, estimates across switch time are relatively stable, and there doesn't appear to be much distinction between the estimates by high exposure threshold.

In figure 2.4 we see more variability in estimates of cumulative death for cancer across switch time regimes than we did for COPD death. For 12 of the 16 switch time regimes

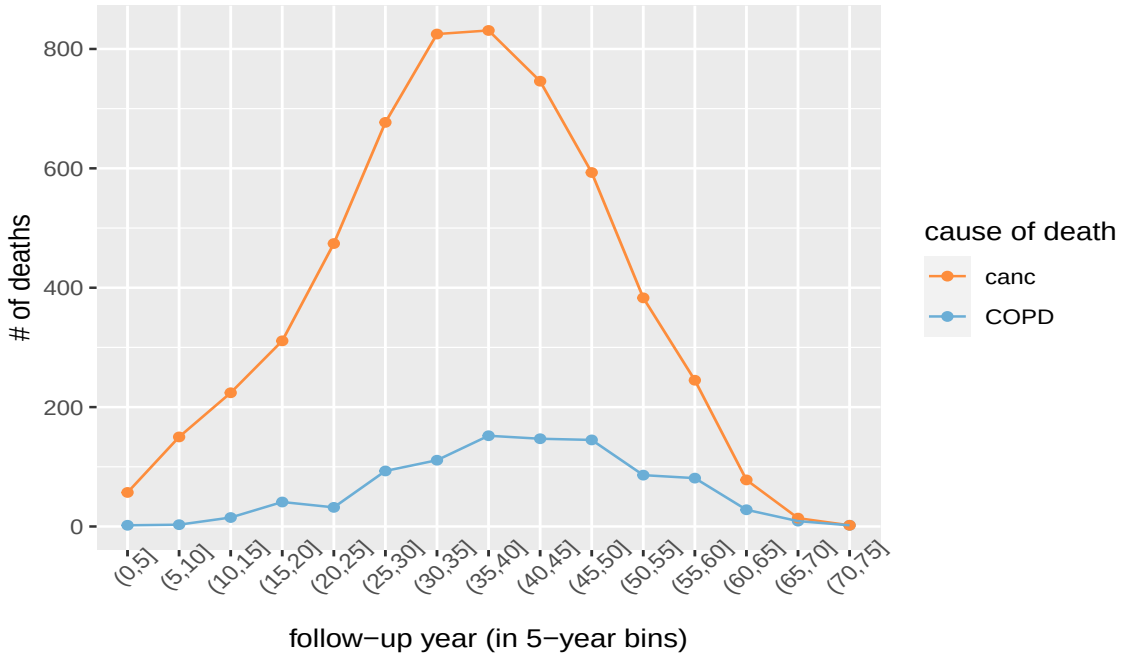

Figure 2.1: Observed number of deaths from COPD and from cancer in the UAW-GM cohort by year of follow-up.

displayed, the estimate using the 90^{th} percentile is lower than that for the 75^{th} percentile.

Risk ratios are presented in table 2.6 which compare the estimated cumulative incidence of death from COPD and from cancer by switch time exposure regime to estimates from the never switch exposure regime. For example, we see that the ratio comparing estimated cumulative incidence of death from COPD if all workers had a cumulative straight MWF exposure above the 75^{th} percentile at the start of follow-up ($S = 0$) relative to the same risk if all workers had a cumulative straight MWF exposure below the 25^{th} percentile throughout follow-up is $0.89(0.86, 0.93)$.

This could be interpreted to mean that a person with cumulative straight MWF exposure over the 75^{th} percentile within the first three years of employment has 89% the risk of dying from COPD within 55 years as someone whose cumulative straight MWF exposure newer exceeds the 25^{th} percentile.

Figures 2.5 and 2.6 show the risk ratios comparing the estimated cumulative incidence of death, from COPD and from cancer, respectively, by switch time exposure regime to estimates from the never switch exposure regime. The example given above is represented by the light blue circle at $(0, 0.89)$ on the plot. Similarly to the estimated cumulative incidences of COPD by switch time regime, we see a large amount of variability in the risk ratios for $S \geq 11$ and an outlier at $S = 11$. The risk ratios are both above and below 1, with

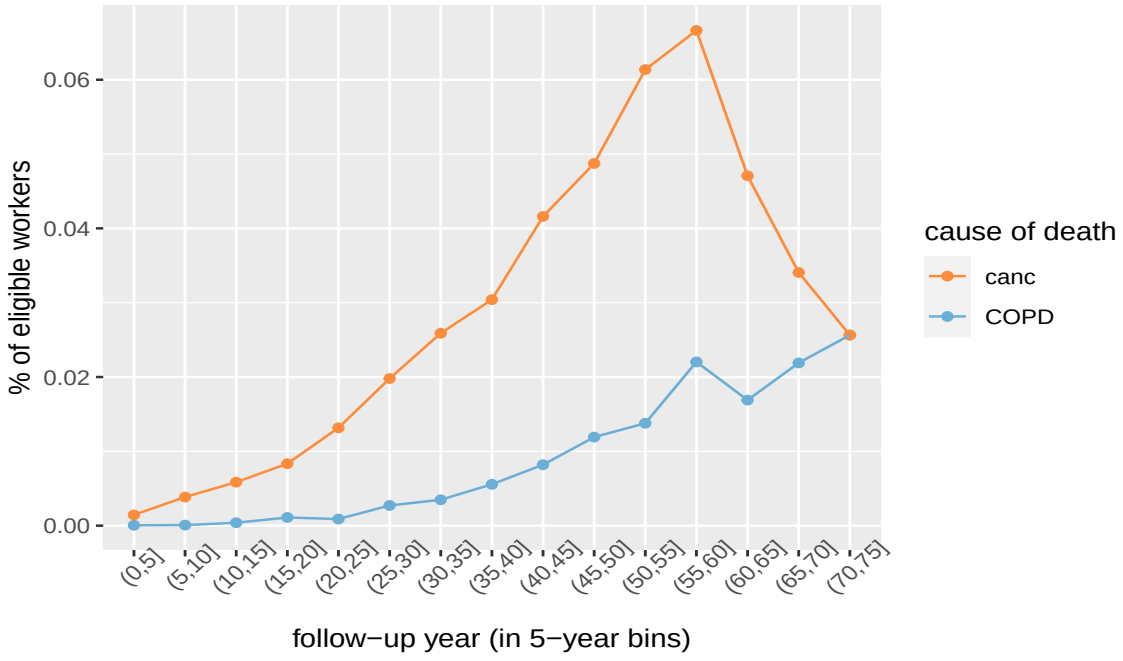

Figure 2.2: Percent of deaths from COPD and from cancer among eligible workers in the UAW-GM cohort by year of follow-up.

no noticeable trend over the values of S or between the estimates using the 75$^{\text{th}}$ and the 90$^{\text{th}}$ percentile.

The risk ratios for cancer deaths (figure 2.6), on the other hand, do seem to be differentiated by the high cumulative exposure threshold used. Counterintuitively, the risk ratios using the 905$^{\text{th}}$ percentile are lower than those using the 75$^{\text{th}}$ percentile, indicating a lower risk of cancer death if workers are forced to have cumulative straight MWF exposure above the 90$^{\text{th}}$ percentile than above the 75$^{\text{th}}$ percentile, relative to exposure below the 25$^{\text{th}}$ percentile. This is possibly explained by the healthy worker effect [18], discussed later.

The cause-specific estimates and 95% confidence intervals of cumulative incidence of COPD death and of cancer death after 55 years of follow-up by switch time (for $S = 0, \ldots, 10$) and by high cumulative straight MWF exposure threshold are presented in table 2.7.These estimates were generated using the cause-specific TMLE estimator, which is identical to the TMLE estimator presented in this paper, except that the competing risk is considered a censoring event. So, for example the cause-specific estimates of the cumulative incidence of COPD death is generated using a censoring variable that includes death from cancer.

These estimates are larger than those generated using the TMLE estimator in 2.5. For example, the cause-specific estimates of the cumulative incidence of death from COPD using the 75$^{\text{th}}$ percentile with switch time intervention regime $S = 1$ is $13.4\%(12.9, 14)$ whereas

Cause of death	**COPD**		**Cancer**		**Total eligible**	
High exposure threshold percentile	**75th**	**90th**	**75th**	**90th**	**75th**	**90th**
Observed switch time	**N (%)**	**N (%)**	**N (%)**	**N (%)**	**N**	
0	64 (3.7)	11 (3.1)	255 (14.8)	62 (17.7)	1720	351
1	17 (3.9)	9 (4)	74 (16.9)	35 (15.4)	438	227
2	12 (2.6)	9 (4.9)	83 (18)	31 (16.8)	460	185
3	12 (3.4)	7 (3.3)	52 (14.8)	33 (15.3)	352	215
4	14 (4.1)	5 (3.7)	52 (15.3)	21 (15.7)	340	134
5	8 (3.1)	2 (1.6)	41 (16.1)	17 (13.7)	254	124
6	5 (2.3)	2 (2.2)	38 (17.8)	15 (16.9)	214	89
7	4 (2)	4 (4.5)	38 (19.2)	18 (20.2)	198	89
8	5 (2.6)	2 (2.3)	38 (19.5)	17 (19.3)	195	88
9	6 (2.9)	3 (3.8)	41 (19.9)	13 (16.5)	206	79
10	3 (1.8)	3 (3.7)	27 (16.2)	11 (13.6)	167	81
11	1 (0.7)	6 (7.9)	17 (12.6)	16 (21.1)	135	76
12	8 (6)	3 (4.6)	21 (15.8)	7 (10.8)	133	65
13	2 (1.8)	2 (3.5)	16 (14.7)	9 (15.8)	109	57
14	4 (4)	1 (1.7)	21 (20.8)	6 (10.3)	101	58
15	0 (0)	0 (0)	0 (0)	0 (0)	89	48
never switch (25th percentile)	492 (2.1%)		3,052 (13.3%)		22,890	

Table 2.4: Observed number of deaths from COPD and from cancer in the UAW-GM cohort by observed switch time regime, and the proportion of COPD and cancer deaths among those eligible.

the TMLE estimate for the same parameter is $9.8\%(7.7, 12)$. There are notable outliers for switch time S = 6, (20.8%), and using the 90th percentile threshold and switch times S = $4, 7, 9$.

Risk ratios comparing the estimated cause-specific cumulative incidence of death from COPD and from cancer by switch time exposure regime to cause-specific estimates from the never switch exposure regime are presented in table 2.8 and visually represented in figures 2.7 and 2.7. These risk ratios are similar to those generated by the TMLE estimator in that they generally vary around 1 with no noticeable trend. The cause-specific estimates for the cumulative incidence of COPD in figure 2.7 for switch times S = $7, 9$ are greater than 2, and seem to stand out among the other estimates.

Cause of Death	**COPD**		**Cancer**	
High exposure threshold percentile	**75th**	**90th**	**75th**	**90th**
Exposure regime switch time (yrs)				
0	6.8 (5.4,8.2)	5.2 (3.8,6.7)	23.2 (20.7,25.6)	28.6 (20.8,26.8)
1	9.8 (7.7,12)	8.9 (7.4,10.4)	24 (21.1,26.9)	28.8 (17.7,22.5)
2	8.6 (7.9,9.4)	8.3 (7.7,8.9)	30.7 (29,32.4)	29.1 (24.2,26.2)
3	5.7 (5,6.4)	9.2 (8.4,10)	24.1 (22.4,25.8)	29.3 (25,27.1)
4	7.5 (6.7,8.4)	7.2 (6.8,7.7)	21.6 (20.5,22.7)	29.5 (18.9,20.4)
5	7.8 (7.4,8.1)	7.6 (7.3,8)	21.8 (20.5,23)	29.7 (16.1,17.9)
6	5.9 (5.3,6.4)	6.7 (6.4,7.1)	25.3 (24.5,26.1)	30 (23.1,24.9)
7	3.5 (3.2,3.7)	11.8 (11.1,12.4)	37.9 (36.7,39.2)	30.2 (26.4,28.3)
8	6.7 (6.1,7.3)	5.8 (5.5,6.1)	28 (26.9,29.1)	30.4 (23.1,24.8)
9	12.7 (11.9,13.6)	6.3 (5.9,6.6)	32.6 (31.6,33.7)	30.6 (23.7,25.5)
10	4.4 (4,4.7)	7.3 (6.8,7.9)	25.5 (24.4,26.7)	30.9 (15.9,18.1)
11	6.1 (5.8,6.4)	34.1 (33.2,35)	20 (19.1,20.9)	31.1 (21.5,23)
12	7.4 (6.8,8)	8.9 (8.4,9.5)	27.7 (19.6,21.1)	31.3 (17.8,19.4)
13	5.8 (5.6,6.1)	13.3 (12.6,13.9)	27.9 (25.8,27.8)	31.5 (22.1,23.9)
14	18.6 (18,19.2)	11.7 (11.1,12.2)	28.1 (24.7,26.7)	31.8 (18.3,19.5)
15	5.3 (5.1,5.6)	5.6 (5.4,5.9)	28.4 (34.8,36.7)	32 (31.1,33)
never switch (25th percentile)	7.6 (6.9,8.4)		24.8 (24,25.7)	

Table 2.5: TMLE estimated cumulative incidence and 95% confidence intervals of death from COPD and from cancer after 55 years of follow-up by switch time and by high cumulative straight metal working fluid exposure threshold.

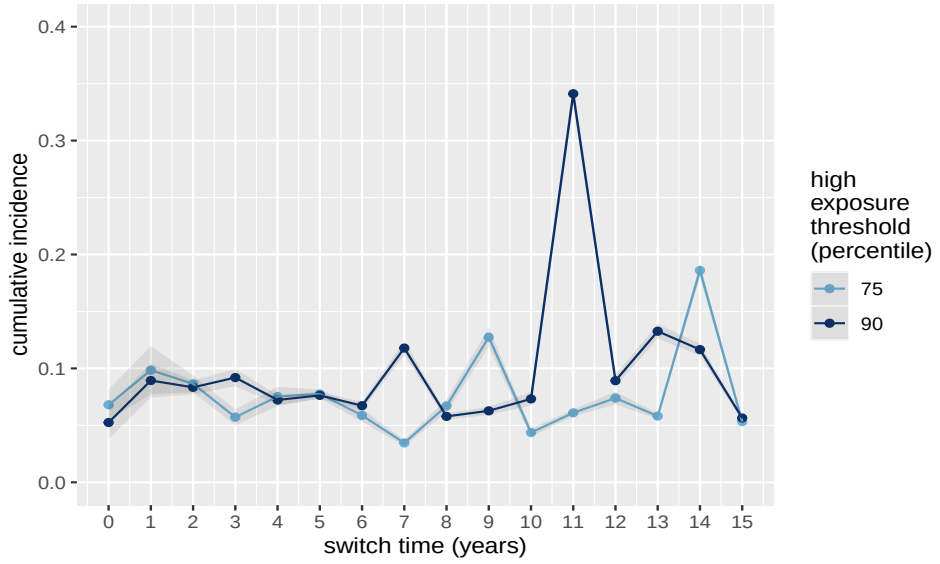

Figure 2.3: TMLE estimated cumulative incidence of death from COPD and 95% confidence intervals after 55 years of follow-up by switch time and high cumulative straight metal working fluid exposure threshold.

2.4 Discussion

In this study we hoped to answer two main research questions. First, to determine if exposure to straight metal working fluid (MWF) causes death from chronic obstructive pulmonary disease (COPD), and secondly to understand the effect using the TMLE estimator described in this paper would have on cumulative incidence estimation. There is little evidence in our analysis to support the hypothesis that straight MWF exposure causes death from COPD. Furthermore, the estimator proposed in this paper generates estimates and confidence intervals very similar to those generated using a standard cause-specific estimator.

In isolation, our lack of evidence to support a causal link between straight MWF exposure and COPD death is not inconsistent with existing literature. However, we also did not uncover a link between MWF exposure and cancer death. There is a lot of strong evidence in existing literature to support a causal link between MWF exposure and cancer death. It might then be possible that there is another aspect of the analysis presented in this paper that explains why there was no link found between MWF exposure and either outcome.

One possible explanation lies in the particular exposure definition used when causally intervening on the data. The exposure metric used is cumulative, meaning that it represents both the intensity and duration of exposure. A person with a particular level of cumulative exposure may have arrived at that exposure by one of two routes: slowly over a long period of time, or with a high exposure over a short period of time. A direct examination of the effect

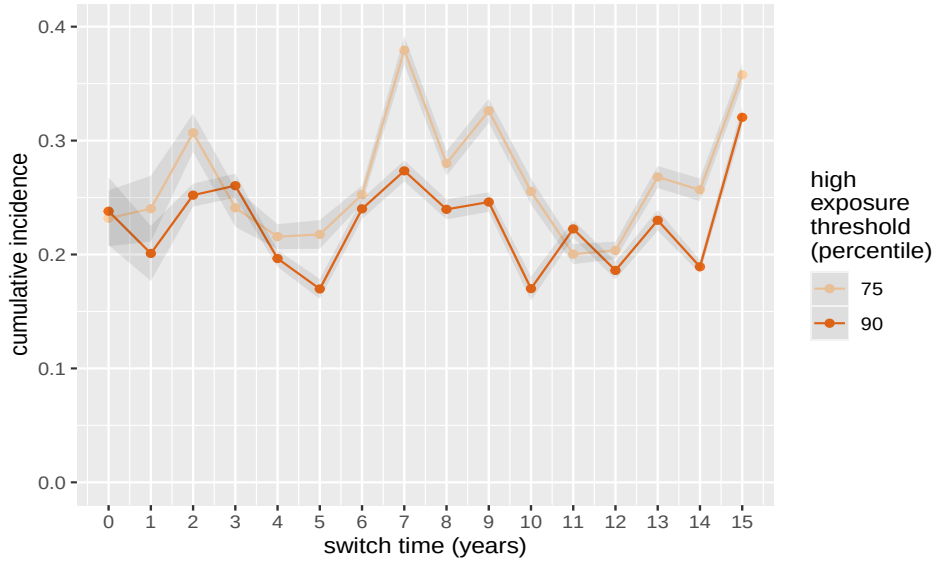

Figure 2.4: TMLE estimated cumulative incidence of death from cancer and 95% confidence intervals after 55 years of follow-up by switch time and high cumulative straight metal working fluid exposure threshold.

of cumulative MWF exposure on COPD risk would not illuminate if there is a differential in risk between these two routes. If such a differential were true, we would want to be able to estimate the effect of cumulative MWF exposure on COPD risk by each route separately, so that the largest effect could be reported.

For this reason, we created the "switch time" model of exposure. Two people with the same level of cumulative exposure but one with an early switch time and one with a later switch time would represent the two routes possible to the same cumulative exposure level. If, for example, a sudden, intense exposure to MWF fluid caused more risk to developing COPD than long exposures (equalling the same cumulative exposure) then we would expect risk estimates for earlier switch times to be higher than for later switch times. The fact that we did not see a distinction between estimates across switch times suggests that such a differential may not exist. However we were only able to examine switch times of 15 or less due to a lack of support over higher switch values. We know, however, that work tenure was on average 18 years, and frequently much longer than that so it could be that a differential would be visible for even later switch times, e.g. greater than 20.

Another shortcoming of this implementation of the "switch time" model of exposure is that it still does not differentiate the ongoing exposure of those with early switch times. A worker whose cumulative exposure crosses the specified threshold in his first year of follow-up and is then subjected to high intensity exposure for the next 10 years is viewed as similarly exposed as a worker whose cumulative exposure crosses the threshold in his first year of

Cause of Death	**COPD**		**Cancer**	
High exposure threshold percentile	**75th**	**90th**	**75th**	**90th**
Exposure regime switch time (yrs)				
0	0.89 (0.86,0.93)	0.69 (0.66,0.72)	0.93 (0.92,0.95)	0.96 (0.94,0.98)
1	1.29 (1.25,1.33)	1.17 (1.14,1.21)	0.97 (0.95,0.99)	0.81 (0.79,0.83)
2	1.13 (1.1,1.17)	1.09 (1.06,1.13)	1.24 (1.22,1.26)	1.02 (1,1.03)
3	0.75 (0.72,0.78)	1.21 (1.17,1.24)	0.97 (0.95,0.99)	1.05 (1.03,1.07)
4	0.99 (0.95,1.02)	0.95 (0.91,0.98)	0.87 (0.85,0.89)	0.79 (0.78,0.81)
5	1.02 (0.98,1.05)	1 (0.97,1.04)	0.88 (0.86,0.89)	0.68 (0.67,0.7)
6	0.77 (0.74,0.8)	0.88 (0.85,0.92)	1.02 (1,1.04)	0.97 (0.95,0.99)
7	0.45 (0.43,0.48)	1.54 (1.5,1.59)	1.53 (1.51,1.55)	1.1 (1.08,1.12)
8	0.88 (0.85,0.91)	0.76 (0.73,0.79)	1.13 (1.11,1.15)	0.97 (0.95,0.98)
9	1.67 (1.63,1.71)	0.82 (0.79,0.85)	1.31 (1.29,1.34)	0.99 (0.97,1.01)
10	0.57 (0.55,0.6)	0.96 (0.93,0.99)	1.03 (1.01,1.05)	0.69 (0.67,0.7)
11	0.8 (0.77,0.83)	4.47 (4.4,4.54)	0.81 (0.79,0.82)	0.9 (0.88,0.91)
12	0.97 (0.94,1)	1.17 (1.13,1.21)	0.82 (0.8,0.84)	0.75 (0.73,0.77)
13	0.76 (0.73,0.79)	1.74 (1.69,1.78)	1.08 (1.06,1.1)	0.93 (0.91,0.94)
14	2.44 (2.39,2.49)	1.53 (1.49,1.57)	1.03 (1.02,1.05)	0.76 (0.75,0.78)
15	0.7 (0.67,0.73)	0.74 (0.71,0.77)	1.44 (1.42,1.46)	1.29 (1.27,1.31)

Table 2.6: Risk ratios comparing TMLE estimated cumulative incidence of death from COPD and from cancer by switch time exposure regimes to estimates from the never switch exposure regime.

follow-up and then accumulates no further exposure. In fact, by treating these two exposure histories the same, our analysis might not be properly adjusting for the healthy worker effect.

We actually do see evidence of the healthy worker effect in the risk ratios of cancer death. The risk ratios for cancer death when using the 75th percentile as the high exposure threshold are higher than those when using the 90th percentile. We would nominally expect estimates using higher thresholds to be higher than those using lower thresholds, so when they are not, it could indicate that the workers with lower exposures had underlying health problems and were therefore were selected out of the jobs with higher exposures. Since we observe this healthy worker effect in the cancer death risk ratios and not in the COPD death risk ratios there could be something deeper that is wrong with this analysis.

Comparing the estimates made using the TMLE estimator defined in this paper to those made using the cause-specific estimator, we see very little difference. There is a slight

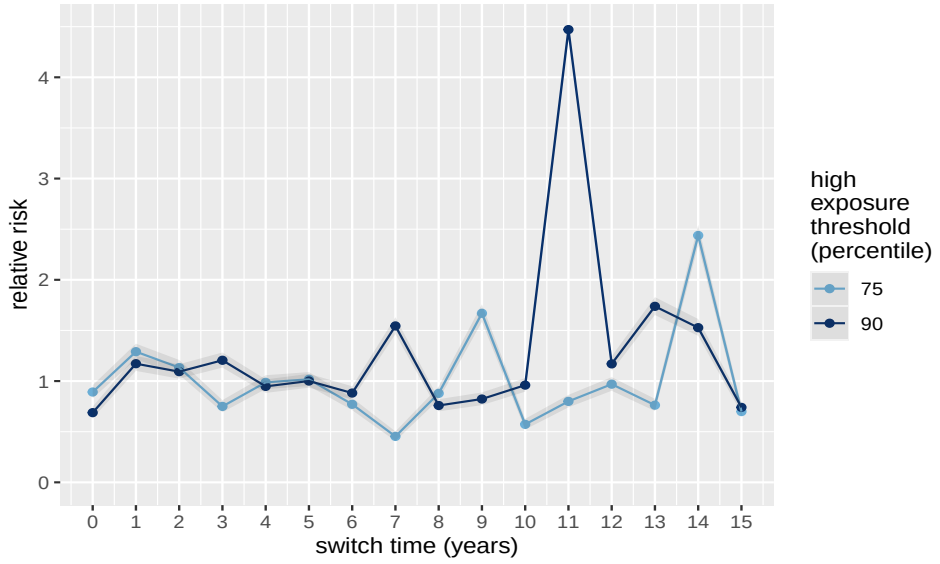

Figure 2.5: Risk ratios comparing TMLE estimated cumulative incidence of death from COPD by switch time exposure regimes to estimates from the never switch exposure regime.

tendency for the cause-specific estimates to vary more between switch times regimes that are close to one another, but this is mild at best.

In conclusion, the switch time model of exposure does not seem to generate clarity regarding the possibly differing effects of straight MWF exposure on COPD and cancer death between those with similar cumulative straight MWF exposure but with contrasting exposure histories. In order to clarify whether there are deeper issues in this analysis, the next step could be to combine the switch time regime estimates into one exposure variable. The results comparing the cumulative incidence of COPD and cancer deaths if all workers switched at anytime and if no workers ever switched could be similar to existing literature. Another possible route would be to use a standard exposure metric such as cumulative exposure or average annual exposure.

Figure 2.6: Risk ratios comparing TMLE estimated cumulative incidence of death from cancer by switch time exposure regimes to estimates from the never switch exposure regime.

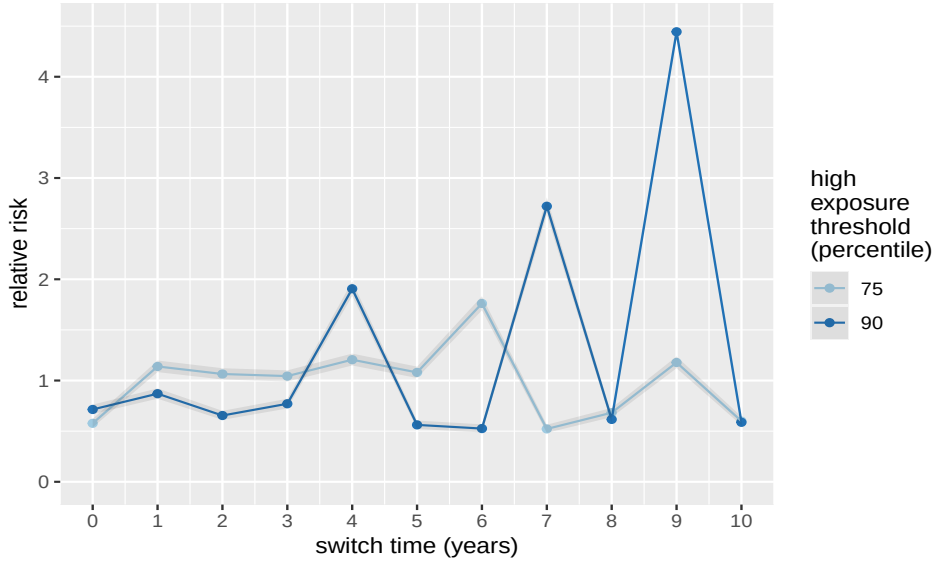

Figure 2.7: Risk ratios comparing cause-specific estimated cumulative incidence of death from COPD by switch time exposure regimes to estimates from the *never switch* exposure regime.

Cause of Death	**COPD**		**Cancer**	
High exposure threshold percentile	**75th**	**90th**	**75th**	**90th**
Exposure regime switch time (yrs)				
0	6.8 (5.1,8.5)	8.4 (7.4,9.4)	22.7 (19.8,25.6)	24.8 (23.1,26.4)
1	13.4 (12.9,14)	10.3 (9.1,11.4)	29.5 (28.9,30.2)	20.3 (19.4,21.3)
2	12.5 (11.6,13.5)	7.7 (6.2,9.2)	33.5 (32.7,34.2)	25.2 (24.1,26.4)
3	12.3 (11.3,13.3)	9.1 (6.8,11.3)	20.4 (19.8,20.9)	33 (32,34)
4	14.2 (12.7,15.7)	22.5 (19.8,25.1)	24.3 (23.5,25)	28.7 (27.9,29.6)
5	12.7 (11.4,14.1)	6.6 (3.7,9.5)	24.4 (23.4,25.3)	17.7 (16.9,18.5)
6	20.8 (19.3,22.2)	6.2 (3.1,9.3)	27.1 (26.5,27.8)	21.6 (20.7,22.5)
7	6.2 (4.7,7.6)	32.1 (28.5,35.6)	39.8 (39,40.7)	27.1 (26.2,28)
8	8 (6.4,9.6)	7.3 (3.8,10.7)	32.6 (31.8,33.5)	22.9 (22.1,23.7)
9	13.9 (12.1,15.7)	52.4 (48.8,55.9)	27.3 (26.4,28.3)	30.5 (29.6,31.4)
10	7.1 (5.3,8.8)	6.9 (6,7.8)	24.8 (23.8,25.7)	16.9 (15.6,18.2)
never switch (25th percentile)	11.8 (11,12.5)		26.4 (25.5,27.2)	

Table 2.7: Estimated cause-specific cumulative incidence and 95% confidence intervals of death from COPD and from cancer after 55 years of follow-up by switch time and by high cumulative straight metal working fluid exposure threshold.

Cause of Death	**COPD**		**Cancer**	
High exposure threshold percentile	**75th**	**90th**	**75th**	**90th**
Exposure regime switch time (yrs)				
0	0.58 (0.54,0.62)	0.71 (0.67,0.76)	0.86 (0.83,0.89)	0.94 (0.91,0.97)
1	1.14 (1.08,1.2)	0.87 (0.82,0.92)	1.12 (1.08,1.16)	0.77 (0.74,0.8)
2	1.06 (1.01,1.12)	0.65 (0.61,0.7)	1.27 (1.23,1.31)	0.96 (0.92,0.99)
3	1.04 (0.99,1.1)	0.77 (0.72,0.82)	0.77 (0.74,0.8)	1.25 (1.21,1.29)
4	1.21 (1.15,1.27)	1.91 (1.84,1.98)	0.92 (0.89,0.95)	1.09 (1.05,1.13)
5	1.08 (1.03,1.14)	0.56 (0.52,0.6)	0.92 (0.89,0.96)	0.67 (0.64,0.7)
6	1.76 (1.69,1.83)	0.53 (0.49,0.57)	1.03 (0.99,1.06)	0.82 (0.79,0.85)
7	0.52 (0.48,0.56)	2.72 (2.64,2.81)	1.51 (1.47,1.55)	1.03 (0.99,1.06)
8	0.68 (0.64,0.73)	0.62 (0.57,0.66)	1.24 (1.2,1.28)	0.87 (0.84,0.9)
9	1.18 (1.12,1.24)	4.44 (4.32,4.57)	1.04 (1,1.07)	1.16 (1.12,1.2)
10	0.6 (0.56,0.64)	0.59 (0.55,0.63)	0.94 (0.91,0.97)	0.64 (0.61,0.67)

Table 2.8: Risk ratios comparing cause-specific estimated cumulative incidence of death from COPD and from cancer by switch time exposure regimes to estimates from the never switch exposure regime.

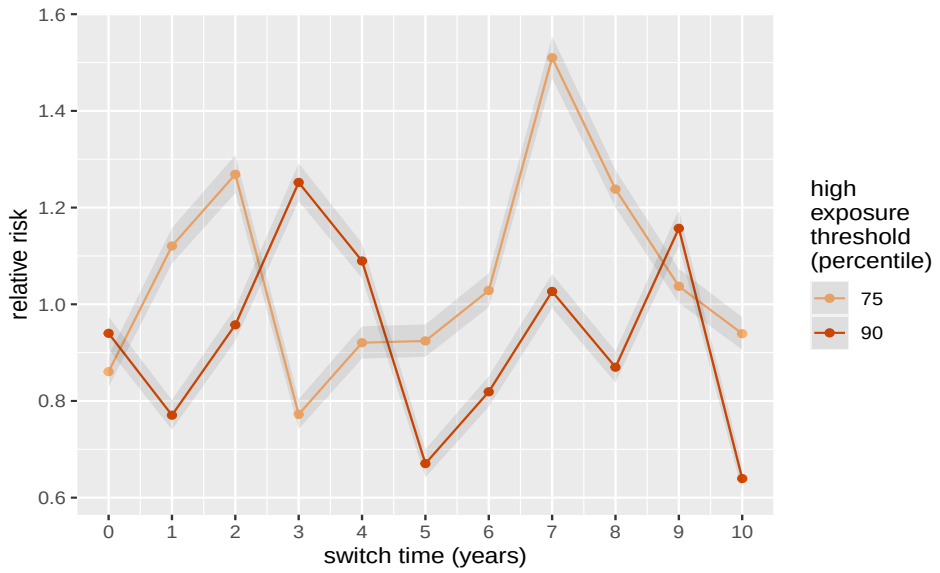

Figure 2.8: Risk ratios comparing cause-specific estimated cumulative incidence of death from cancer by switch time exposure regimes to estimates from the never switch exposure regime.

Chapter 3

Simulation Comparing Methods of Competing Risk Treatment in Estimation

3.1 Introduction

In a longitudinal study in which a researcher is interested in the time from study entry until a particular event, subjects may experience the event of interest, or one of many other multiple risks. These outcomes are called competing risks when, by their nature, observing one prohibits the observing of another. For example, in an occupational epidemiology study workers may be followed in order to study the effects of exposure to fine particulate matter on cardiovascular disease. However, workers most affected by exposure may develop sub-clinical disease and leave their job before the outcome of interest develops. Current approaches to adjust for competing risks tend to have complicated interpretations, or impose unrealistic assumptions upon the relationship among the potential outcomes.

As a motivating example, consider a group of autoworkers observed over many years in order to study the effects of factory exposure to metal working fluids (MWF) on mortality due to chronic obstructive pulmonary disease (COPD). A worker's vital status is followed until time of death, which may be attributed to one of multiple competing causes, or administrative censoring at end of follow-up. COPD takes a long time to develop, during which time workers may die from other diseases with shorter latent periods, including one of several specific cancers that have been associated with MWF exposure. In order to compare the cumulative incidence of COPD after a certain number of years of followup and under specified exposure regimes, the estimation techniques used by researchers typically involve treating the competing risk as a censoring event, i.e. imposing an assumption of independence upon the two outcomes to employ classic survival estimation techniques.

Kaplan and Meier introduced the product-limit estimator to estimate the cause-specific survival in a theoretical experiment in which all causes of leaving a cohort are eliminated

except the one of interest. The authors stipulate the assumption necessary for their estimator to be unbiased as independence between the cause of interest and the other competing causes. The product- limit estimator non-parametrically estimates the *cause-specific survival function*, that is, the proportion, $S(t)$, of the cohort whose lifetime (or time until the event of interest) exceeds t in the absence of competing risks without placing assumptions on the form of $S(t)$. The Nelson-Aalen estimator is the corresponding estimator of the cumulative incidence function, $\Lambda(t)$ and its properties have been analyzed elsewhere [33, 1].

In this simulation study we generate an occupational cohort with work-related exposures until their death or end of observation. We examine bias from competing risks and estimator selection by varying the dependence of outcomes on exposure. We compare estimate bias across estimator, and consider its effects on interpretation.

We focus on two estimators to addressing competing risks: treating the competing risk as a censoring event (i.e. the cause-specific estimator) and treating the competing risk as a time-varying covariate. In both scenarios, we will be using the causal roadmap to estimate the parameters corresponding to the entire outcome distribution including the outcome of interest and the competing risk. We propose wider use of the second scenario which does not rely on the assumption of sequential independence between the outcome and the competing risk.

3.2 Methods

The simulation study addresses the performance of the TMLE estimator under two settings, in which the competing risk is treated either as a

Setting 1: time-varying covariate

Setting 2: censoring event

It will be noted below when the analyses under each setting diverge.

3.2.1 Observed data structure

Consider a longitudinal study of autoworkers with outcome COPD death and exposure defined as cumulative straight metal working fluid switching from low to high, as defined by a pre-specified percentile. Observations are taken at baseline, three intervention nodes $t = 0, 1, 2$, and the outcome is observed at time $K = 3$.

Time-varying variables are generated for each time point and include a covariate representing time off work, L, cumulative exposure, A, censoring, C, and outcomes, Y_1 and Y_2. We arbitrarily select Y_1 as the outcome of interest and Y_2 as the competing risk, as the analysis is the same in either case.

W : baseline categorical variable representing 6 combinations of race and plant

$A(0)$: cumulative exposure to straight MWF in first three years of working

At times $t = 1, 2$

$L(t)$: time-varying time off work

$Y_1(t)$: death by Chronic Obstructive Pulmonary Disease (COPD)
$\in$ {0=no death from COPD, 1=death from COPD}

$Y_2(t)$: death by any cancer $\in$ {0=no death from cancer, 1=death from cancer}

$A(t)$: time-varying cumulative exposure to straight MWF $\in$ {0=low, 1=high}

$C(t)$: censoring $\in$ {0=not censored, 1=censored}
before time t by one of the following events:

- administrative censoring
- observed age exceeds 108
- death due to an unrelated cause, i.e. causes other than cancer or COPD

At time $K = 3$ we observe the final outcomes, $Y_1(3)$ and $Y_2(3)$.

3.2.2 Structural Causal Model

We assume after baseline covariate, W, and baseline cumulative exposure, $A(0)$, the following time ordering at times $t = 1, 2, 3$:

$$W \to L(t), Y_1(t), Y_2(t) \to A(t) \to C(t)$$

Let $U = (U_W, U_L, U_{Y_1}, U_{Y_2}, U_A, U_C)$ be unobserved exogenous variables.
We define the following structural equations using $f = (f_W, f_{A(t)}, f_{L(t)}, f_{C(t)}, f_{Y_1(t)}, f_{Y_2(t)})$

such that for $t = 1, 2, 3$:

$$
\begin{aligned}
W &= f_W(U_W) \\
A(0) &= f_{A_0}(W, U_A) \\
L(1) &= f_{L(1)}(W, A(0), U_L) \\
Y_1(1) &= 0 \text{ (outcomes } Y_1 \text{ and } Y_2 \text{ are not possible at first time point)} \\
Y_2(1) &= 0 \\
A(1) &= f_{A(1)}(W, A(0), U_A) \\
C(1) &= f_{C(1)}(W, U_C) \\
L(2) &= f_{L(2)}(W, A(1), L(1), C(1), U_L) \\
Y_1(2) &= f_{Y_1(2)}(W, A(1), L(2), C(1), U_{Y_1}) \\
Y_2(2) &= f_{Y_2(2)}(W, A(1), L(2), C(1), U_{Y_2}) \\
A(2) &= f_{A(2)}(W, A(1), L(2), Y_1(2), Y_2(2), C(1), U_A) \\
C(2) &= f_{C(2)}(W, Y_1(2), Y_2(2), C(1), U_C) \\
L(3) &= f_{L(3)}(W, A(2), L(2), Y_1(2), Y_2(2), C(1), U_L) \\
Y_1(3) &= f_{Y_1(3)}(W, A(2), L(3), Y_1(2), Y_2(2), C(2), U_{Y_1}) \\
Y_2(3) &= f_{Y_2(3)}(W, A(2), L(3), Y_1(3), Y_2(2), C(2), U_{Y_1})
\end{aligned}
$$

Then the observed data are n i.i.d. draws of $O = (W, A(0), L(1), Y_1(1), Y_2(1), A(1), C(1), L(2), Y_1(2), Y_2(2), A(2), C(2), Y_1(3), Y_2(3)) \sim P_0 \in \mathcal{M}^F$, where $\mathcal{M}^F$ is the structural causal model containing all possible distributions of P_0.

Under the setting in which the competing risk is treated as a censoring event, the censoring variable, C, is defined to include instances of the competing risk, Y_2. As such, Y_2 is omitted from the structural equations, and the parent nodes of censoring events are altered to include the parent nodes of the competing risk. We write the following to define the

pre-intervention structural equations.

$$
\begin{aligned}
W &= f_W(U_W) \\
A(0) &= f_{A_0}(W, U_A) \\
L(1) &= f_{L(1)}(W, A(0), U_L) \\
Y_1(1) &= 0 \text{ (outcome is not possible at first time point)} \\
A(1) &= f_{A(1)}(W, A(0), U_A) \\
C(1) &= f_{C(1)}(W, U_C) \\
L(2) &= f_{L(2)}(W, A(1), L(1), C(1), U_L) \\
Y_1(2) &= f_{Y_1(2)}(W, A(1), L(2), C(1), U_{Y_1}) \\
A(2) &= f_{A(2)}(W, A(1), L(2), Y_1(2), C(1), U_A) \\
C(2) &= f_{C(2)}(W, A(1), L(2), Y_1(2), C(1), U_C) \\
L(3) &= f_{L(3)}(W, A(2), L(2), Y_1(2), C(2), U_L) \\
Y_1(3) &= f_{Y_1(3)}(W, A(2), L(3), Y_1(2), C(2), U_{Y_1})
\end{aligned}
$$

Then the observed data are n i.i.d. draws of $O = (W, A(0), L(1), Y_1(1), A(1), C(1), L(2), Y_1(2), A(2), C(2), Y_1(3)) \sim P_0 \in \mathcal{M}^F$, where $\mathcal{M}^F$ is the structural causal model containing all possible distributions of P_0.

3.2.3 Interventions and Counterfactual Outcomes

Under the first setting, where the competing risk is considered a time-varying confounder, we consider the static exposure regimes A_s for $s = 0, 1, 2, 3$ in which the exposure level is set to 0 until a specified "switch" time, s, where exposure switches to 1 and remains 1 until end of observation. We can call A_0 "always high" and A_3 "never high". Let $C_0 = (C(1) = 0, C(2) = 0)$ indicate the censoring history with no censoring. We are interested in comparing the cumulative incidence of COPD death after 3 time points across switch time regimes so for convenience we suppress the censoring notation and define the static exposure and censoring interventions as follows.

$$
\begin{aligned}
A_0 &= (1, 1, 1) \\
A_1 &= (0, 1, 1) \\
A_2 &= (0, 0, 1) \\
A_3 &= (0, 0, 0)
\end{aligned}
$$

Let $Y_1^{A_s}(3)$ be the (possibly contrary to fact) counterfactual value of $Y_1(3)$ for an individual under treatment regime A_s. We intervene on our structural equations to generate a post-intervention distribution which include the counterfactual outcomes.

For example, for the exposure regime $A_1 = (0, 1, 1)$, the post-intervention structural equations are:

$$
\begin{aligned}
W &= f_W(U_W) \\
A(0) &= 0 \\
L(1) &= f_{L(1)}(W, A(0) = 0, U_L) \\
Y_1(1) &= 0 \\
Y_2(1) &= 0 \\
A(1) &= 1 \\
C(1) &= 0 \\
L(2) &= f_{L(2)}(W, A(1) = 1, L(1), C(1) = 0, U_L) \\
Y_1(2) &= f_{Y_1(2)}(W, A(1) = 1, L(2), C(1) = 0, U_{Y_1}) \\
Y_2(2) &= f_{Y_2(2)}(W, A(1) = 1, L(2), C(1) = 0, U_{Y_2}) \\
A(2) &= 1 \\
C(2) &= 0 \\
L(3) &= f_{L(3)}(W, A(2) = 1, L(2), Y_1(2), Y_2(2), C(1) = 0, U_L) \\
Y_1^{A_s}(3) &= f_{Y_1(3)}(W, A(2) = 1, L(3), Y_1(2), Y_2(2), C(2) = 0, U_{Y_1}) \\
Y_2^{A_s}(3) &= f_{Y_2(3)}(W, A(2) = 1, L(3), Y_1(3), Y_2(2), C(2) = 0, U_{Y_1})
\end{aligned}
$$

Under the second setting where the competing risk is considered a censoring event, the interventions and counterfactual are annotated the same as they were in the previous setting. The definition of C has changed, but the notation has not.

The post-intervention structural equations under this setting omit Y_2 so, for example, for

the exposure regime $A_1 = (0, 1, 1)$ the structural equations look like:

$$\begin{aligned}
W &= f_W(U_W) \\
A(0) &= 0 \\
L(1) &= f_{L(1)}(W, A(0) = 0, U_L) \\
Y_1(1) &= 0 \\
A(1) &= 1 \\
C(1) &= 0 \\
L(2) &= f_{L(2)}(W, A(1) = 1, L(1), C(1) = 0, U_L) \\
Y_1(2) &= f_{Y_1(2)}(W, A(1) = 1, L(2), C(1) = 0, U_{Y_1}) \\
A(2) &= 1 \\
C(2) &= 0 \\
L(3) &= f_{L(3)}(W, A(2) = 1, L(2), Y_1(2), C(1) = 0, U_L) \\
Y_1(3) &= f_{Y_1(3)}(W, A(2) = 1, L(3), Y_1(2), C(2) = 0, U_{Y_1})
\end{aligned}$$

3.2.4 Causal Quantity of Interest

The causal quantity of interest is the mean outcome of $Y_1(3)$ under the interventions of interest.

$$\Psi^F(P_0) = \mathbb{E}_{P_0}(Y_1^{A_s}(3))$$

for $s = 0, 1, 2, 3$. We can calculate causal contrasts, e.g. the causal risk difference of the cumulative incidence of COPD death between exposure intervention regimes A_0, *always high* and A_3, *never high* as $RD = E(Y_1^{A_0}(3)) - E(Y_1^{A_3}(3))$.

The notation for the causal quantity of interest is identical between the setting in which the competing risk is treated as a time-varying covariate and the one in which it is treated as a censoring event.

3.2.5 Statistical Estimand and Identfication

The causal quantity of interest is a parameter of the distributions of counterfactuals $Y_1^{A_s}(3)$ for $s = 0, 1, 2, 3$. To estimate this quantity from the observed data, we need to determine that we can express the distribution of $Y_1^{A_s}(3)$ in terms of P_0, the distribution of the observed data.

We must make two additional assumptions to allow us to identify the distribution of $Y_1^{A_s}(3)$ [41]. If these additional assumptions to the full structural causal model (SCM) are true, then there exists a function, ψ, such that $\Psi^F(P_0) = \psi(P_0)$. We call the resulting full SCM $\mathcal{M}^{F*} \subset \mathcal{M}^F$, i.e. $\mathcal{M}^{F*}$ is the structural statistical model containing all distributions P_0 such that

$$\psi(P) = \Psi^F(P_0)$$

for P the distribution of O implied by P_0. Now, $O \sim P_0 \in \mathcal{M}$. If we believe the following assumptions, then $P_0 \in \mathcal{M}^*$.

The sequential randomization assumption means that at each time point, all common causes of L, Y and A are observed. The positivity assumption means that each observation has a positive probability of following the intervention regime of interest at each time point.

Sequential Randomization Assumption

$$A(0) \perp Y_1^{A_s}(3)|W$$

$$A(1) \perp Y_1^{A_s}(3)|W, A(0)$$

$$A(2) \perp Y_1^{A_s}(3)|W, A(1), L(2), Y_2(2), C(1)$$

Positivity Assumption

$$P_0(A(0) = 1|W) > 0$$

$$P_0(A(1) = 1|W, A(0) > 0$$

$$P_0(A(2) = 1|W, A(0), L(2) > 0$$

Versions of these assumptions must also hold for censoring which is to say that all common causes of L, Y, and C are observed, and each observation has a positive probability of not being censored at each time. These assumptions look the same under both competing risks settings, except for the modified definition of censoring to include Y_2, explained below as an additional sequential randomization assumption.

Let Q_{Y_1}, Q_L, Q_W be the conditional distributions of $Y_1(3)$, $L(t)$ for $t = 1, 2, 3$ and W, and let g_A, g_C be the conditional distributions of A and C, given their histories. Please note that the distributions of $Y_1(1)$ and $Y_1(2)$ are included in that of $L(1)$ and $L(2)$.

Under our first setting where $Y_2(t)$ is treated as a time varying covariate, the distribution of $Y_2(t)$ can be included in that of $L(t)$ for $t = 1, 2, 3$. Then the observed data likelihood can be factorized as follows:

$$\begin{aligned} P(O) =& Q_W(W) \times Q_{Y_1}(Y_1(3)|W, A(2), L(3), C(2)) \\ & \times Q_L(L(3)|W, A(2), L(2), C(2)) \\ & \times Q_L(L(2)|W, A(1), L(1), C(1)) \\ & \times Q_L(L(1)|W, A(0)) \\ & \times g_C(C(2)|W, C(1)) \\ & \times g_C(C(1)|W) \\ & \times g_A(A(2)|W, A(1), L(2), C(1)) \\ & \times g_C(A(1)|W, A(0)) \end{aligned}$$

Under the second setting where $Y_2(t)$ is treated as a censoring event, the distribution of $Y_2(t)$ can be included in that of $C(t)$ for $t = 1, 2$. Then the observed data likelihood can be factorized as follows:

$$\begin{aligned} P(O) =& Q_W(W) \times Q_{Y_1}(Y_1(3)|W, A(2), L(3), C(2)) \\ & \times Q_L(L(3)|W, A(2), L(2), C(2)) \\ & \times Q_L(L(2)|W, A(1), L(1), C(1)) \\ & \times Q_L(L(1)|W, A(0)) \\ & \times g_C(C(2)|W, A(1), L(2), C(1)) \\ & \times g_C(C(1)|W) \\ & \times g_A(A(2)|W, A(1), L(2), C(1)) \\ & \times g_C(A(1)|W, A(0)) \end{aligned}$$

Under these assumptions, our causal quantity of interest, $\Psi^F(P_0) = \mathbb{E}_{P_0}(Y_1^{A_s}(3))$, can be identified by setting the intervention variables according to A_s and then taking a sequence of recursively defined conditional expectations called the G-computation formula[7].

$$\begin{aligned} \Psi^F(P_0) =& \mathbb{E}_{P_0}(Y_1^{A_s}(3)z) \\ =& \sum_w \{ \sum_l [P_0(Y_1^{A_s}(3)|W = w, L(3) = l_3, (A, C) = A_s) \\ & \times P_0(L(3) = l_3|W = w, L(2) = l_2, (A, C) = A_s) \\ & \times P_0(L(2) = l_2|W = w, L(1) = l_1, (A, C) = A_s) \\ & \times P_0(L(1) = l_1|W = w, (A, C) = A_s)] \\ & \times P_0(W = w)\} \\ =& \sum_w \{ \sum_l [P_0(Y_1(3) = 1|W = w, L(3) = l, (A, C) = A_s) \\ & \times Q_L(L(3) = l_3|W = w, L(2) = l_2, (A, C) = A_s) \\ & \times Q_L(L(2) = l_2|W = w, L(1) = l_1, (A, C) = A_s) \\ & \times Q_L(L(1) = l_1|W = w, (A, C) = A_s)] \\ & \times Q_W(W = w)\} \\ =& \psi(P_0) \end{aligned}$$

where the right hand side is expressed entirely with P_0 with the intervention variables set to A_s with probability 1 and is our statistical estimand, $\psi(P_0)$. The statistical estimand under the second setting where the competing risk is treated as a censoring event is expressed the same as above, but has the alternative definition of censoring to include Y_2. As such, in this second setting there is an additional sequential randomization assumption placed on Y_2 because it is included as a censoring event.

Additional Sequential Randomization Assumption (under setting where competing risk is treated as a censoring event)

$$Y_2(2) \perp Y_1^{A_s}(3) | W, A(1), L(2), C(1)$$

3.2.6 Simulated data

We simulate data to mimic an occupational cohort where $10,000$ workers are observed at baseline and then at three following time points. A baseline covariate, W, is distributed as a multinomial random variable with six categories which represent combinations of race, sex, and plant as distributed in the observed data. The first cumulative exposure, variable, A_0, is also determined at baseline. Time-varying variables are generated for each time point and include a covariate representing time off work, L, the outcome of death from COPD, Y_1, the competing risk of death from any cancer, Y_2, cumulative exposure, A, and censoring, C. These variables are then generated according to the following rules.

$$\begin{aligned}
W &\sim \text{multinom}(p_1 = 0.11, p_2 = 0.29, p_3 = 0.16, p_4 = 0.14, p_5 = 0.30) \\
A(0)|W &\sim \text{expit}(-2 - 0.45W) \\
L(1)|W, A(0) &\sim \text{expit}(0.5 + 0.25W + A(0)) \\
Y_1(1) &= 0 \\
Y_2(1) &= 0 \\
A(1)| \quad W, A(0), Y_1(1) = 0, Y_2(1) = 0, &\sim \min(1, A(0) + \text{expit}(-0.5 + 0.25W5A(0))) \\
C(1)| \quad W, A(1), Y_1(1) = 0, Y_2(1) = 0 &\sim \text{expit}(-1.5 - 0.1W + 2A(1)) \\
L(2)| \quad W, L(1), A(1), Y_1(1) = 0, Y_2(1) = 0, C(1) = 0 &\sim \text{expit}(0.5 - 0.25W + A(1) + 0.5L(1)) \\
Y_1(2)| \quad W, A(1), L(2), Y_1(1) = 0, Y_2(1) = 0, C(1) = 0 &\sim \text{expit}(\beta_1 + W + \alpha_1 A(1) - 0.25L(2)) \\
Y_2(2)| \quad W, A(1), L(2), Y_1(2) = 0, Y_2(1) = 0, C(1) = 0 &\sim \text{expit}(\beta_1 + W + \alpha_1 A(1) - 0.25L(2))
\end{aligned}$$

$$
\begin{array}{lll}
A(2)| & W, A(1), L(2), & \sim \min(1, A(1) + \text{expit}(-0.5 + 0.25W + A(1) + 0.5L(2))) \\
 & Y_1(2) = 0, Y_2(2) = 0 & \\
 & C(1) = 0 & \\
C(2)| & W, A(2), L(2) & \sim \text{expit}(-1.5 - 0.1W + 2A(2) + L(2)) \\
 & Y_1(2) = 0, Y_2(2) = 0 & \\
 & C(1) = 0 & \\
L(3)| & W, A(2), L(2) & \sim \text{expit}(0.5 + 0.25W2A(2) + 0.5L(2)) \\
 & Y_1(2) = 0, Y_2(2) = 0, & \\
 & C(2) = 0 & \\
Y_1(3)| & W, L(3), A(2) & \sim \text{expit}(\beta_2 + W + \alpha_2 A(2) - 0.25L(3)) \\
 & Y_1(2) = 0, Y_2(2) = 0, & \\
 & C(2) = 0 & \\
Y_2(3)| & W, L(3), A(2) & \sim \text{expit}(\beta_2 + W + \alpha_2 A(2) - 0.25L(3)) \\
 & Y_1(3) = 0, Y_2(2) = 0, & \\
 & C(2) = 0 &
\end{array}
$$

Since we are generating our own data, we know the true Q values, that is, the conditional distributions of L and Y_1 (and Y_2 in the setting where Y_2 is considered a time-varying covariate). To evaluate our true parameter, we simulate draws from Q, but intervene on g, that is, the conditional distributions of A and C (and Y_2 in the setting where Y_2 is considered a censoring event). We intervene to set C (and Y_2) to not occur (no censoring) and we set $A(0), A(1), A(2)$ according to one of the interventions of interest, A_s for $s = 0, 1, 2, 3$. We calculate the statistical estimand using this data by taking the mean of Y_3. Under sequential randomization and the positivity assumption, this is the causal quantity of interest.

3.2.7 Estimators

Using the ltmle R package [31], we use the TMLE estimator to estimate the cumulative incidence of the outcome, Y_1, under both settings where the competing risk is treated as a time-varying covariate and where it is treated as a censoring event.

For each of n observations we denote the expectation of Y conditional on the full history of observation i which followed the exposure regime A_s and was fully observed as $\overline{Q}_i(3) = \mathbb{E}(Y_i|\overline{L}_i(3), Y_{1,i}(2) = 0, Y_{2,i}(2) = 0, A_s, C_i(2) = 0, W_i)$. Then, $\overline{Q}_i(2) = \mathbb{E}(\overline{Q}_i(3)|\overline{L}_i(1), Y_{1,i}(1) = 0, Y_{2,i}(1) = 0, A_s, C_i(t) = 0, W_i)$ where $\overline{Q}_i(1) = \mathbb{E}(\overline{Q}_i(2)|W_i)$. Then, the G-computation estimator is defined as follows.

$$\hat{\psi}_{\text{GComp}} = \frac{1}{n}\sum_{i=1}^{n}\overline{Q}_i(1)$$

The TMLE estimator is implemented by updating the above conditional outcome for each subject, $\overline{Q}_i(t)$ with $\overline{Q}_i^*(t)$ such that for the clever covariate $G(t)(C(t), A_s, \overline{L}(t-1), W) = \mathbb{I}(C(t) = 0, A_s)$. We calculate $\overline{Q}_i^*(t) = \text{expit}(\text{logit}(\overline{Q}_i(t) + \hat{\epsilon}_i(t)G_i(t)(A_s, C(t) = 0, \overline{L}(t-1))$

where $\hat{\epsilon}(t)$ is the fitted coefficient of the above clever covariate. Then, the TMLE estimator is as follows.

$$\hat{\psi}_{\text{TMLE}} = \frac{1}{n}\sum_{i=1}^{n}\overline{Q}_i^*(1)$$

We use the above estimators under the first competing risk setting, where we treat the competing risk, Y_2, as a time-varying covariate. After this event occurs, the remaining distributions become degenerate. There is also a version of these estimators for the setting in which we categorize the competing risk as a censoring event, to be intervened upon.

3.2.8 Simulations

We simulated data to mimic a longitudinal occupational cohort study where workers are observed at baseline and for three time periods thereafter or until they experience one of two outcomes, death from COPD or death from cancer, or they are censored.

We generated 100 datasets of cohorts containing $n = 10,000$ workers. At baseline, a covariate predictive of censoring and both outcomes of interest are generated, and at each time point, time-varying variables are generated that include a covariate representing time off work, cumulative exposure, censoring, and outcomes.

Recall that the cumulative exposure variable can be interpreted as the cumulative straight metal working fluid switching from low to high, as defined by a pre-selected percentile. Our estimates are generated by intervening on the cumulative exposure histories with interventions where exposure level is set as low until a specified switch time, where exposure switches to high and remains high until end of observation. So the exposures of interest can be described as $A_0 = (1,1,1)$, "always high", $A_1 = (0,1,1)$, "switch low to high at t=1", $A_2 = (0,0,1)$, "switch low to high at t=2", and $A_3 = (0,0,0)$, "never high".

We also intervene to preclude censoring. In the second setting where we consider the competing risk a censoring event, we also intervene to set $Y_2(1) = 0$ and $Y_2(2) = 0$.

Across our simulations, we vary the strength of the effect of cumulative exposure on COPD and cancer death as parameterized by the coefficients in the structural equations generating the outcomes, $\alpha = (\alpha_1, \alpha_2)$ across $\alpha = \{(3,3),(1,1),(1,3),(3,1)\}$.

We compare the difference in cumulative incidence of COPD death and of cancer death under causal treatment regimes A_0, *always high* and A_3, *never high*. In both competing risk settings we use the TMLE estimator, but in the setting where we treat the competing risk as a censoring event, we call our estimator "cause-specific". All analyses use the R ltmle package [31] with Super Learner [47] with a library that included simple mean, generalized linear modeling, and a stepwise procedure. We also employ naive estimators by calculating the conditional means of the outcome given the workers were uncensored and followed the specified exposure regime, ignoring all other relationships among the variables.

All estimates are compared to the true risk difference, as calculated by the mean of the binary outcome variable, $Y_1(3)$, across a single large-sample ($n = 1,000,000$) simulation with exposure and censoring mechanisms specified per the *always high* and *never high* interventions. Since the sequential randomization assumption holds given the structural equations

described in this paper, the TMLE estimator is consistent with the causal effect, and this method is a valid way to calculate the true risk difference.

3.3 Results

		COPD death			
Estimator	$\boldsymbol{\alpha_1, \alpha_2}$	**RD**	**bias**	**SE**	**MSE**
TMLE	(1,1)	0.049	0	0.011	<0.001
	(1,3)	0.04	-0.001	0.01	<0.001
	(3,3)	0.359	0.005	0.016	<0.001
	(3,1)	0.391	-0.001	0.015	<0.001
cs TMLE	(1,1)	0.052	0.002	0.01	<0.001
	(1,3)	0.052	0.011	0.011	<0.001
	(3,3)	0.4	0.046	0.019	0.002
	(3,1)	0.404	0.012	0.016	<0.001
naive	(1,1)	0.023	-0.027	0.007	<0.001
	(1,3)	0.014	-0.027	0.006	<0.001
	(3,3)	0.167	-0.187	0.015	0.035
	(3,1)	0.209	-0.183	0.013	0.034
cs naive	(1,1)	0.025	-0.024	0.007	<0.001
	(1,3)	0.024	-0.016	0.008	<0.001
	(3,3)	0.217	-0.136	0.018	0.019
	(3,1)	0.223	-0.17	0.016	0.029

Table 3.1: Simulation risk difference estimates of COPD death between *always high* and *never high* intervention regimes, and under four coefficient parameter values, α.

Comparing the TMLE and the cause-specific TMLE estimator performances in tables 3.1, and 3.2 we see a lower bias, standard error (SE), and mean squared error (MSE) for the TMLE estimator across values of α. The bias for the TMLE estimator is on the order of the standard error for the true risk difference ($1/\sqrt{n}$ for $n = 1,000,000$), demonstrating the consistency of that estimator. Both TMLE estimators outperform the naive estimators.

Recall that α is the coefficient in the structural equation used to simulate the data that reflects the strength of the effect of cumulative exposure on COPD and cancer death. All estimators report a larger risk difference for larger α relative to the outcome of interest. For example, in table 3.1 we see the largest risk differences when $\alpha_1 = 3$, and in table 3.2 we see the largest risk differences when $\alpha_2 = 3$. We also see the largest bias values for $\alpha = 3$ indicating that the consistency of the estimators is challenged as the effect of the exposure on either the outcome or the competing risk increases.

		Cancer death			
Estimator	α_1, α_2	**RD**	**bias**	**SE**	**MSE**
TMLE	(1,1)	0.047	0	0.01	<0.001
	(1,3)	0.376	0.002	0.016	<0.001
	(3,3)	0.251	-0.001	0.014	<0.001
	(3,1)	0.019	-0.002	0.009	<0.001
cs TMLE	(1,1)	0.049	0.002	0.01	<0.001
	(1,3)	0.387	0.012	0.019	<0.001
	(3,3)	0.302	0.05	0.017	0.003
	(3,1)	0.03	0.008	0.009	<0.001
naive	(1,1)	0.022	-0.025	0.007	<0.001
	(1,3)	0.199	-0.175	0.014	0.031
	(3,3)	0.107	-0.145	0.012	0.021
	(3,1)	0.006	-0.016	0.005	<0.001
cs naive	(1,1)	0.023	-0.024	0.008	<0.001
	(1,3)	0.211	-0.163	0.017	0.027
	(3,3)	0.159	-0.093	0.015	0.009
	(3,1)	0.014	-0.008	0.006	<0.001

Table 3.2: Simulation risk difference estimates of cancer death between *always high* and *never high* intervention regimes, and under four coefficient parameter values, α.

Figures 3.1 and 3.2 illustrate the consistency of the longitudinal TMLE estimator when the competing risk is treated as a time-varying covariate.

3.4 Discussion

It is challenging to estimate outcomes in longitudinal studies with competing risks. The findings in this paper demonstrate that treating a competing risk as a time-varying covariate is a valid method of consistent estimation. When using the causal roadmap [48], we can identify the expected counterfactual outcome when the structural models generate the competing risk as a time-varying covariate by correctly classifying the competing risk in the g-computation formula. This method relies on the sequential randomization of exposure and censoring, and on the positivity assumption.

When we repeat the causal roadmap with structural models that generate the competing risk as a censoring event, we can similarly identify the expected counterfactual outcome by correctly classifying the competing risk in the g-computation formula. However, this method relies on an additional assumption, that sequential randomization holds for the competing risk.

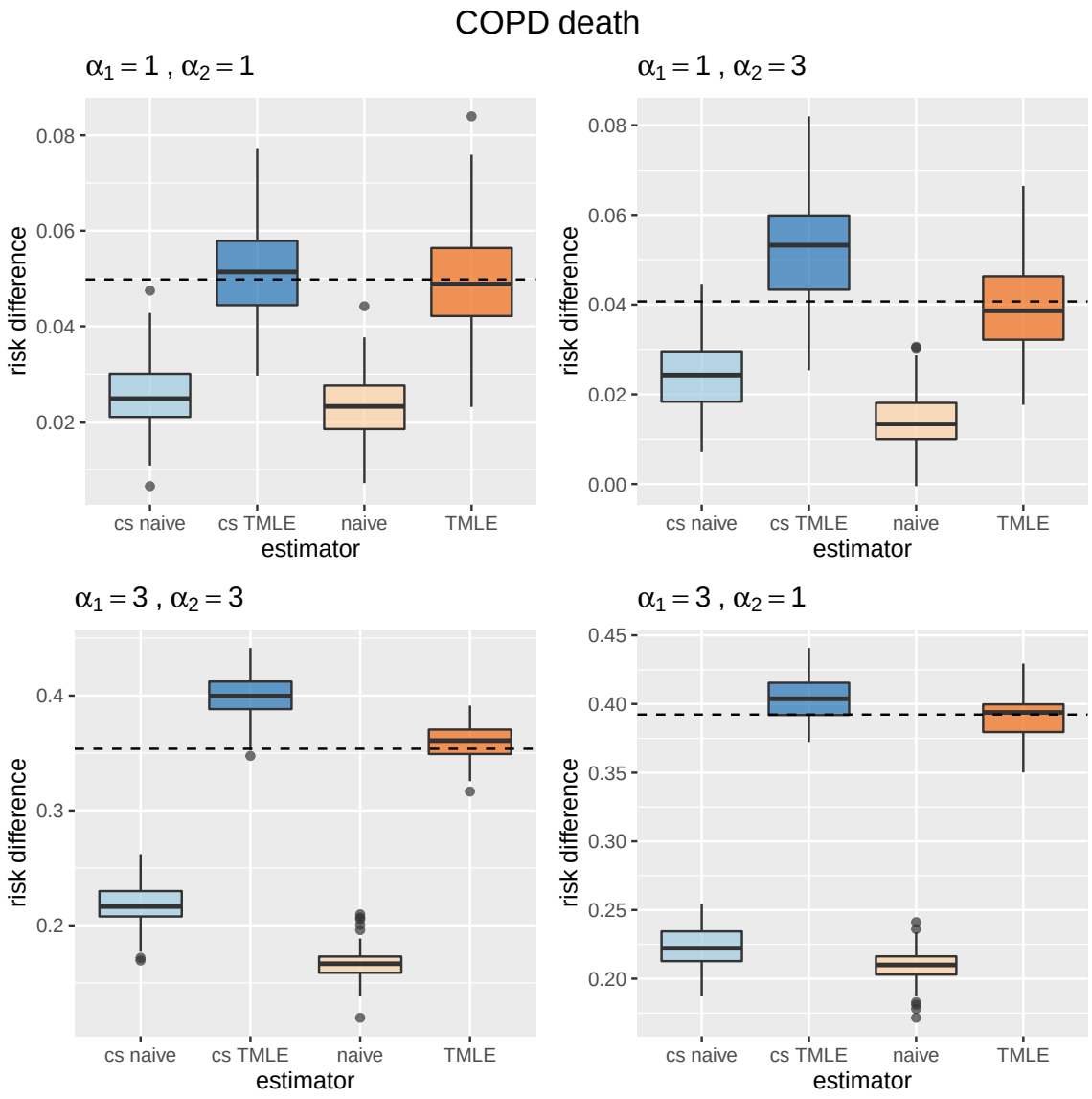

Figure 3.1: Boxplots of the simulation risk difference estimates of COPD death compared to the truth (dashed line) under the *always high* and *never high* interventions, and under four coefficient parameter values, α_1, α_2.

In an occupational cohort study, it is natural to suspect that exposure to harmful chemicals at work can affect multiple competing disease processes over a lifetime. However, there may be unmeasured factors that affect the risk of death for multiple causes of death. If such a factor exists, then when applying the causal roadmap to estimate the risk of a particular cause of death, if we treat a competing cause of death as a censoring event, it is not possible

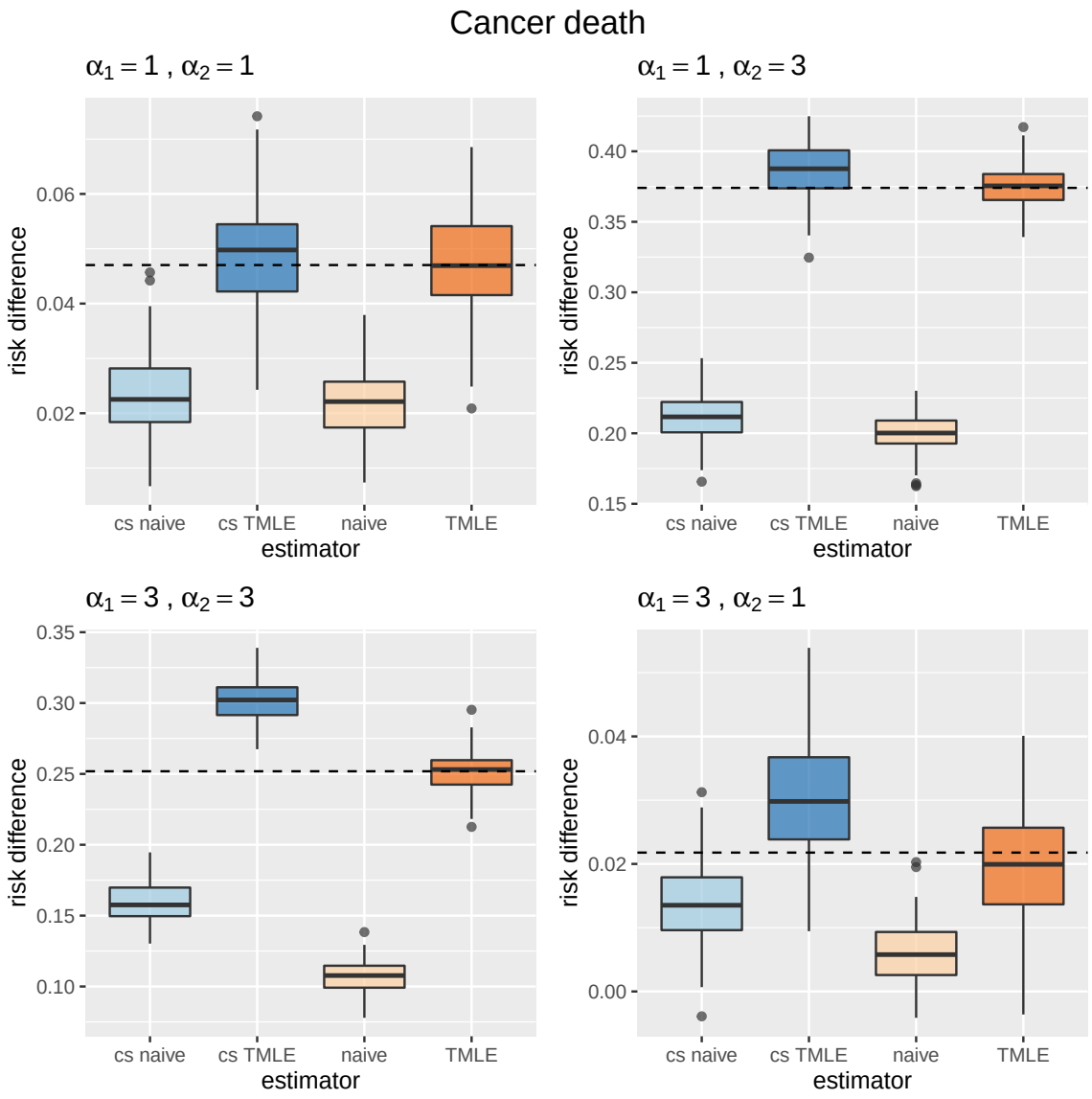

Figure 3.2: Boxplots of the simulation risk difference estimates of cancer death compared to the truth (dashed line) under the *always high* and *never high* interventions, and under four coefficient parameter values, α_1, α_2.

to identify the causal quantity with a statistical estimand based on the observed data. To do so would require the sequential randomization assumption to hold for the competing risk, but in fact we could not adjust for all common causes of the outcome of interest and the competing risk.

However, this paper demonstrates that there exists an alternative method of estimation

which does not rely on adjusting for unmeasured variables. When classifying a competing risk as a time-varying covariate, the sequential randomization assumption does not need to apply to the competing risk. When using an existing unbiased estimator such as TMLE, the resulting estimates are consistent with the true causal quantity, and maintain the existing properties of that estimator.

Bibliography

[1] Odd Aalen. "Nonparametric inference for a family of counting processes". In: *The Annals of Statistics* (1978), pp. 701–726.

[2] Odd O Aalen and Søren Johansen. "An empirical transition matrix for non-homogeneous Markov chains based on censored observations". In: *Scandinavian Journal of Statistics* (1978), pp. 141–150.

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

[4] PK Anderson et al. "Statistical models based on counting processes". In: *Biometrics* 24 (1993), pp. 100–101.

[5] MD Attfield and K Morring. "An investigation into the relationship between coal workers' pneumoconiosis and dust exposure in U.S. coal miners". In: *American Industrial Hygiene Association Journal* 53.8 (1992), pp. 486–492.

[6] Peter C Austin and Jason P Fine. "Practical recommendations for reporting F ine-G ray model analyses for competing risk data". In: *Statistics in medicine* 36.27 (2017), pp. 4391–4400.

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

[8] Joseph Berkson and Lila Elveback. "Competing exponential risks, with particular reference to the study of smoking and lung cancer". In: *Journal of the American Statistical Association* 55.291 (1960), pp. 415–428.

[9] Norman E Breslow. "Contribution to discussion of paper by DR Cox". In: *J. Roy. Statist. Soc., Ser. B* 34 (1972), pp. 216–217.

[10] 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: application in a cohort of autoworkers exposed to metalworking fluids". In: *Epidemiology* (2012), pp. 212–219.

[11] Stephen R Cole and Miguel A Hernán. "Constructing inverse probability weights for marginal structural models". In: *American journal of epidemiology* 168.6 (2008), pp. 656–664.

[12] Stephen R Cole et al. "Estimation of the standardized risk difference and ratio in a competing risks framework: application to injection drug use and progression to AIDS after initiation of antiretroviral therapy". In: *American journal of epidemiology* 181.4 (2015), pp. 238–245.

[13] Jerome Cornfield. "Estimation of the probability of developing a disease in the presence of competing risks". In: *American Journal of Public Health and the Nations Health* 47.5 (1957), pp. 601–607.

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

[15] David R Cox. "Partial likelihood". In: *Biometrika* 62.2 (1975), pp. 269–276.

[16] David R Cox. "Regression models and life-tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202.

[17] KT Drzewiecki et al. "Clinical course of cutaneous malignant melanoma related to histopathological criteria of primary tumour". In: *Scandinavian journal of plastic and reconstructive surgery* 14.3 (1980), pp. 229–234.

[18] Ellen A Eisen, Sally Picciotto, and James M Robins. "Healthy worker effect". In: *Wiley StatsRef: Statistics Reference Online* (2014).

[19] 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.

[20] Jason P Fine and Robert J Gray. "A proportional hazards model for the subdistribution of a competing risk". In: *Journal of the American statistical association* 94.446 (1999), pp. 496–509.

[21] Richard D Gill, Mark J Van Der Laan, and James M Robins. "Coarsening at random: Characterizations, conjectures, counter-examples". In: *Proceedings of the First Seattle Symposium in Biostatistics*. Springer. 1997, pp. 255–294.

[22] Robert J Gray. "A class of K-sample tests for comparing the cumulative incidence of a competing risk". In: *The Annals of statistics* (1988), pp. 1141–1154.

[23] Sander Greenland and James M Robins. "Identifiability, exchangeability, and epidemiological confounding". In: *International journal of epidemiology* 15.3 (1986), pp. 413–419.

[24] MF 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.

[25] E Cuyler Hammond and Daniel Horn. "The relationship between human smoking habits and death rates: a follow-up study of 187,766 men". In: *Journal of the American Medical Association* 155.15 (1954), pp. 1316–1328.

[26] Daniel F Heitjan and Donald B Rubin. "Ignorability and coarse data". In: *The annals of statistics* (1991), pp. 2244–2253.

[27] Edward L Kaplan and Paul Meier. "Nonparametric estimation from incomplete observations". In: *Journal of the American statistical association* 53.282 (1958), pp. 457–481.

[28] 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).

[29] Deborah D Landen et al. "Coal dust exposure and mortality from ischemic heart disease among a cohort of US coal miners". In: *American journal of industrial medicine* 54.10 (2011), pp. 727–733.

[30] Bryan Lau, Stephen R Cole, and Stephen J Gange. "Competing risk regression models for epidemiologic data". In: *American journal of epidemiology* 170.2 (2009), pp. 244–256.

[31] Samuel D Lendle et al. "ltmle: an R package implementing targeted minimum loss-based estimation for longitudinal data". In: *Journal of Statistical Software* 81.1 (2017), pp. 1–21.

[32] DY Lin. "On the Breslow estimator". In: *Lifetime data analysis* 13.4 (2007), pp. 471–480.

[33] Wayne Nelson. "Hazard plotting for incomplete failure data". In: *Journal of Quality Technology* 1.1 (1969), pp. 27–52.

[34] Jerzy Neyman. *First Course in Probability and Statistics*. 1950.

[35] Judea Pearl. *Causality*. Cambridge university press, 2009.

[36] Maya L Petersen and Mark J van der Laan. "Causal models and learning from data: integrating causal modeling and statistical estimation". In: *Epidemiology (Cambridge, Mass.)* 25.3 (2014), p. 418.

[37] 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* (2014), pp. 436–443.

[38] Ross L Prentice et al. "The analysis of failure times in the presence of competing risks". In: *Biometrics* (1978), pp. 541–554.

[39] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: `https://www.R-project.org/`.

[40] International Agency for Research on Cancer, International Agency for Research on Cancer, et al. *Overall evaluations of carcinogenicity: an updating of IARC Monographs volumes 1 to 42*. IARC Lyon, France: 1987.

[41] 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.

[42] James M Robins, Miguel Angel Hernan, and Babette Brumback. *Marginal structural models and causal inference in epidemiology*. 2000.

[43] James M Robins and Andrea Rotnitzky. "Recovery of information and adjustment for dependent censoring using surrogate markers". In: *AIDS epidemiology*. Springer, 1992, pp. 297–331.

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

[45] R Sapir-Pichhadze et al. "Survival analysis in the presence of competing risks: the example of waitlisted kidney transplant candidates". In: *American Journal of Transplantation* 16.7 (2016), pp. 1958–1966.

[46] Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." In: *Statistical Science* (1990), pp. 465–472.

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

[48] Mark J Van der Laan and Sherri Rose. *Targeted learning in data science*. Springer, 2018.

[49] Mark J Van Der Laan and Daniel Rubin. "Targeted maximum likelihood learning". In: *The international journal of biostatistics* 2.1 (2006).

[50] Marcel Wolbers et al. "Prognostic models with competing risks: methods and application to coronary risk prediction". In: *Epidemiology* (2009), pp. 555–561.