Title
Applications of Causal Inference to Problems of Occupational Epidemiology

Author
Brown, Daniel Martin

Publication Date
2014

Peer reviewed|Thesis/dissertation

Applications of Causal Inference to Problems of Occupational Epidemiology

by

Daniel Martin Brown

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor in Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Maya Petersen, Co-Chair
Professor Mark J. van der Laan, Co-Chair
Professor Jennifer Ahern
Professor Ellen A. Eisen

Spring 2014

Applications of Causal Inference to Problems of Occupational Epidemiology

Abstracts

Applications of Causal Inference to Problems of Occupational Epidemiology

by

Daniel Martin Brown

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Maya Petersen, Co-Chair
Professor Mark J. van der Laan, Co-Chair

This dissertation concerns the application of the techniques of causal inference to problems of occupational health. The abstracts of the three works which comprise the primary substance of this dissertation are reproduced below.

The healthy worker survivor effect (HWSE) is a feature of occupational cohort studies which can lead to biased estimates of the etiologic effects of exposures if the estimation procedure does not account for its sources. The HWSE arises from underlying temporal processes characteristic of working populations in which time-varying health status is a criteria for entry into follow-up as well as both a predictor and a consequence of exposure. We distinguish two sources of HWSE: left-truncation in the presence of heterogeneous susceptibility as well as time-varying confounding on the causal pathway. We apply longitudinal minimum-loss-based estimation to simulated data in order to illustrate the effect of each process on estimates of exposure response, and clarify the extent to which methodological solutions can properly adjust for the bias.

We consider the problem of the estimation of parameters of the full-data distribution from data structures in which some confounding variables are unmeasured in a portion of the population. Our focus is on evaluating approaches to implementation of an augmented inverse probability of censoring weighted targeted minimum-loss based estimator (A-IPCW TMLE) first proposed by Rose and van der Laan (2012). This is an inverse probability weighted estimator in which estimation proceeds using a reweighted set of fully observed data points. The weights used are the inverses the estimated probability of being fully observed which is then augmented by an estimate of the expectation of the full data influence function, given the always observed variables. The estimator's performance is compared to standard weighting approaches and multiple imputation in both a simulation study and an applied data example.

We investigate the effect of cumulative exposure to particulate matter with an aerodynamic diameter <2.5 μm ($PM_{2.5}$) on the incidence of ischemic heart disease (IHD) in a cohort

of aluminum workers followed for 15 years, adjusting for time-varying confounding affected by prior exposure. We use longitudinal targeted minimum-loss based estimation (TMLE) to estimate the cumulative risk difference for IHD if always exposed above an exposure cut-off compared to always exposed below, while never censored. We stratify our analyses by sub-cohort employed in the smelters versus fabrication facilities. We selected two exposure cut-offs *a priori*, at the median and 10th percentile, within each sub-cohort. In smelters, the estimated IHD risk difference after 15 years is 2.1% (-1.3%, 5.5%) if always exposed compared to never exposed above the median cut-off of 1.77 $\frac{mg}{m^3}$ and 2.9% (0.6%, 5.1%) using the 10th percentile cutoff of 0.10 $\frac{mg}{m^3}$. For fabrication workers, the estimated risk difference is 0.9% (-1.6%, 3.5%) using the median cut-off of 0.20 $\frac{mg}{m^3}$ and 2.5% (0.8%, 4.1%) using the 10th percentile cut-off of 0.06 $\frac{mg}{m^3}$. Results are presented as marginal incidence curves, describing the cumulative risk of IHD for each sub-cohort under each intervention regimen. By control of the time-varying confounding on the causal pathway that characterizes healthy worker survivor effect, TMLE estimated associations between cumulative $PM_{2.5}$ exposure and IHD that were not detectable using standard analytical techniques in a previous report. This represents the first longitudinal application of TMLE, a method for generating doubly robust semi-parametric efficient substitution estimators, in the field of occupational and environmental epidemiology.

This dissertation is dedicated to the lights of my life, my wife Gwendy, and my daughter, Solana.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank,

Ellen Eisen for her tireless support throughout this process. I am endlessly appreciative for the time and energy that she has put in to teaching me and working with me, also for her enthusiasm, her honesty and her kindness. I also want to acknowledge the Alcoa and epidemiologic research teams here in Environmental Health: Sadie Costello, Sally Piccioto, Jonathan Chevrier, Andreas Neophytou, Betsey Noth, Kathy Hammond, and Liza Lutzker. Thank you all for the hundreds of hours of stimulating conversations that taught me so much.

The chairs of my committee, Mark and Maya. Through their efforts, I have learned to hold myself up to a higher standard of work and thought. I aspire to conduct my professional life in as rigorous and principled fashion as they have. Thank you both also for your patience and kindness in guiding me through this process. Thank you to Jen for your participation, for your insightful comments, and for inspiring me to take the next steps to stay in academia.

The Centers for Occupational and Environmental Health for providing the financial support for my education, and for introducing me to a field in which I could apply the tools I learned. I would not have believed that occupational health would be my passion, but as I progress in my education, I am continuously inspired by it. Work forms the center of our lives, and to do so in environments that are not only safe and healthy but also bring sustenance and meaning should be a fundamental right for everyone.

My fellow students, particularly Luca Pozzi, Molly Davies, Sam Lendle, and Stephanie Sapp. Thank you for being friends and resources and for showing up for reading review sessions.

My family, my mother and father Gary and Ginger, my brothers Geoff and Sandro, my sisters Sam and Meredith, who have always been there for me. Thank you all for being the foundation of my life and for inspiring me with your work and energy and love.

My friends who have also been kind, supportive, and a source of welcome respite from my cares.

Josy Lismay, who has been an invaluable source of wisdom and compassion.

My wife Gwendy, who has been and will always be my best friend, my confidant and my love. Thank you for being my rock when I needed support and my wind when I needed to fly. Thank you also to my daughter Solana Mae for being so full of love and light. The days I spend with you are the best of them.

Chapter 1

Introduction

This dissertation several issues common to occupational cohort studies through the lens of causal inference. We demonstrate the ability of estimation methods motivated by the use of causal frameworks to correct for time-varying confounding, selection bias, and health-dependent censoring. We also highlight remaining barriers to unbiased estimation of causal parameters commonly arising in occupational studies. We also explore an approach previously proposed by Rose and van der Laan (2012) for the application of targeted minimum-loss based estimation in data structures with missing confounders, whose finite sample performance had not yet been compared to common alternatives. We finish with an applied example that suggests a causal relationship between heart disease and airborne particulate exposure in a population of aluminum workers.

Occupational cohort studies are generally performed with the intent of estimating the causal relationship between a workplace exposure and morbidity of the cohort. The healthy worker effect (HWE) is a feature of occupational studies that must be accounted for during the estimation process, as it can lead to inaccurate estimates of this true response if the natural correlation between employment and health is not accounted for. This correlation occurs due to two overarching effects. The first is the healthy hire effect (HHE), which occurs because healthy people are more likely to seek employment and to be offered jobs, especially the strenuous and dangerous jobs often related to the exposures under study. The second reason is the healthy worker survivor effect (HWSE), in which healthy workers are more likely to stay at work and therefore both accrue more exposure and have a larger probability of being included into the cohort. The HWSE is a special case of time-varying confounding on the causal pathway (Arrighi and Hertz-Picciotto, 1994) in which health status serves as the mediator of the effects of exposure on both disease and work termination.

In some cases the HWE can manifest due to a selection process in which health status serves as a determinant of participation in a study cohort (Hernán et al., 2005). This selection process often results in data structures with non-ignorable (Rubin, 1976a) missing data

patterns, via the following pathway. It is the case that at any point in time, the population of active workers will contain a relatively healthy subset of the worker cohort. It is also often true that researchers require additional information on potential confounders, which are not regularly collected as part of the administrative databases that form the backbone of the observed data. This information therefore often gathered on a convenient subset of the entire cohort, usually consisting of a cross-sectional sample of active workers. The observed data then contains a missing data problem where the assumption of no unmeasured confounding is only true among a relatively healthy subset of the population. We address both of these issues from a theoretical perspective in the first two chapters of this work, and the example which completes the dissertation synthesizes this work in an applied setting.

In chapter two, we present the results of a simulation study that highlights several aspects of occupational cohort studies. We first demonstrate the effects of time varying confounding on the causal pathway on standard estimators as well as the ability of causally motivated estimators to adjust for this confounding. We then examine the effects of the combination of left-truncation and susceptibility heterogeneity on the effect estimates derived from prevalent cohorts. Finally we demonstrate the use of longitudinal targeted minimum loss-based estimation (LTMLE) to estimate causal parameters when worker follow-up continues past employment termination. We formalize the effects we are observing using the language of directed acyclic graphs (DAGs) (Pearl, 1995, Greenland et al., 1999a) and argue for a broader understanding of the healthy worker survivor effect that includes both time varying confounding and left truncation in combination with heterogeneous susceptibility.

In chapter three, we consider the problem of the estimation of parameters of the full-data distribution from data structures in which some confounding variables are unmeasured in a portion of the population. Our focus is on evaluating approaches to implementation of an augmented inverse probability of censoring weighted targeted minimum-loss based estimation (A-IPCW TMLE) first proposed by Rose and van der Laan (2012). This is an inverse probability weighted estimator (Li et al., 2011) in which estimation proceeds using a reweighted set of fully observed data points. The weights used are inverses of an estimate of the probability of being fully observed which is then augmented by an estimate of the expectation of the full data influence function, given the always observed variables. The estimator's performance is compared to standard weighting approaches and multiple imputation in both a simulation study and an applied data example.

In chapter four, we present an applied example of an analysis of the relationship between occupational exposure to $PM_{2.5}$ and incidence of ischemic heart disease. We apply LTMLE to a cohort of aluminum smelter and fabrication workers in order to estimate the causal effect of exposure to airborne particulate matter with an aerodynamic diameter $<2.5\ \mu m$ on the incidence of ischemic heart disease. We present our results in the form of adjusted survival curves predicting the estimated cumulative incidence of heart disease among all workers had they been continually exposed above and below exposure cut-offs while staying at work until retirement. We also provide a detailed walkthrough of the steps undergone to

create these estimators and a summary of the properties of the LTMLE estimator, designed for an epidemiologic audience.

Appendix A contains a literature review compiling some of the key contributers to the understanding of the healthy worker survivor effect and some concluding remarks.

Chapter 2

Simulating the Healthy Worker Survivor Effect

2.1 Introduction

Recent approaches to dealing with the healthy worker survivor effect in occupational epidemiology (Chevrier et al., 2012a, Dumas et al., 2013a, Naimi et al., 2014, Picciotto et al., 2014) have focused on the problem of time varying confounding affected by prior exposure. This occurs when time-varying health status serves as both a mediator between occupational exposure and future disease as well as a predictor of future exposure. Decline in time-varying health status will reduce future exposure, as workers in poorer health will tend to reduce their hours, change to a lower exposed job, or to leave work altogether. Standard methods of confounding adjustment will not generate unbiased estimators when analyzing data from such studies, as health status at time t is both a confounder of exposure response at future times $t+$ and a mediator of the effect of exposure at earlier times $t-$. Several methodological solutions to this problem have been suggested, including G-estimation of a structural nested model (Robins, 1987, Hernán et al., 2005), inverse probability of treatment weighted (IPTW) estimation of the parameters of a marginal structural model (Robins, 2000b, 1999, Hernán et al., 2000), and targeted minimum loss-based estimation (TMLE)(van der Laan and Gruber, 2012, Bang and Robins, 2005, Stitelman Ori et al., 2012, Schnitzer et al., 2013). All of these approaches can be loosely classified as causal methods.

Applications of these methods to occupational cohorts have stressed their ability to generate unbiased estimates in the presence of time-varying confounding and have concluded that they have corrected for the HWSE (Chevrier et al., 2012a, Dumas et al., 2013a, Naimi et al., 2014). While we agree with these statements, there are additional aspects of the HWSE that are not addressed by this approach. These missing aspects are the effects of

left-truncation and selection bias in the presence of heterogeneity of susceptibility. They can occur when an occupational cohort is a prevalent cohort, comprising of workers who were hired prior to the start of follow-up (Brookmeyer et al., 1987, Wang et al., 1993, Cole et al., 2004, Howards et al., 2006). Such workers have been subject to both the exposure and the risk of the disease of interest prior to the start of follow-up, and these may affect their probabilities of remaining at work until cohort follow-up starts. In contrast, an incident cohort of workers whose follow-up starts immediately at hire has not been subjected to the same selection pressures.

Within this working population there may be a range of susceptibility to the exposure and the disease of interest. We distinguish between two possible ways in which this susceptibility could function in the population. We define exposure susceptibility as an effect modifier of the exposure-disease relationship. We define disease susceptibility as increasing disease risk independently from exposure.

If either type of susceptibility is heterogeneous in the population, a prevalent cohort will tend to have fewer susceptible workers than an incident cohort of new hires. This is due in part to left truncation, as those workers with the shortest survival times (less than the time between hire and follow-up start) will not be at risk for incident disease and therefore not be eligible for cohort inclusion. Similarly, workers who leave work in this interim time between hire and follow-up start will not be available for cohort inclusion. These phenomona can alternatively be viewed as a form of selection bias (Hernán et al., 2004). Estimation in the prevalent cohort involves conditioning on both active work status and the absence of prior disease, which could be consequences of both the exposure and disease of interest (Cole et al., 2010). We will revisit these points later to make a distinction between left truncation and selection bias.

The result of these phenomena is an observed population with a different mix of baseline characteristics, including past exposure history, than would be observed in the cohort in which all workers are followed from hire. The effect of exposure as estimated in the prevalent cohort may then not be generalizable to the incident cohort. If the question of interest is the effect of exposure among the population of all workers, rather than a subset of survivors, left truncated cohorts in the presence of susceptibility heterogeneity present another source of health worker survivor bias. This bias was explored previously in a simulation study by Applebaum, Malloy and Eisen, who concluded that the bias due to left truncation and heterogeneous susceptibility operated in the absence of the HWSE as defined by time varying confounding (Applebaum et al., 2011).

Another salient feature of occupational studies is that follow-up may continue past employment termination (Picciotto et al., 2013). As unemployed workers can generally not be exposed, the data structure contain subject-times in which the probability of exposure is 0. This is a violation of a positivity assumption commonly made to give a causal interpretation to estimates generated by some methods. This has been used as a justification for

the exclusive use of G-estimation of a structural nested model in occupational epidemiology (Naimi et al., 2013, Robins, 2000b), as the causal interpretation of these parameters do not rely on this assumption. There are alternative parameters for defining the causal effect that do not rely on this assumption and can be estimated in data structures with follow-up past employment termination (see for example the review in Petersen et al. (2012)). To our knowledge, these methods have not been applied in the field of occupational epidemiology in either simulated or applied examples.

2.2 Data, Models and Simulation

2.2.1 Data Description

We consider the following full data structures X and observed data structures for incident (O_i) and prevalent (O_p) cohorts, where:

$$X = (W, S, A(1), L(1), \ldots, A(K), L(K))$$

$$O_i = (W, A(1), L(1), \ldots, A(K), L(K))$$

$$O_p = (W, A(1), L(1), \ldots, A(c), L(c), \ldots A(c+K), L(c+K) | Y(c-1) = 0, C(c-1) = 1),$$

where W is a vector of measured baseline variables that may be confounders of the relationship between $A(t)$ and $L(t)$ and S is an unmeasured indicator of susceptibility to the exposure and/or disease of interest. $A(t)$ is the treatment of interest and contains two nodes: $E(t)$ is a binary exposure experienced by a worker during year t and $C(t)$ is an indicator of active work status at the end of time point t, so $C(t) = 1$ indicates that a worker was actively employed. $L(t) = (H(t), Y(t))$ are time-varying covariates measured at the end of year t. $H(t)$ is an indicator that a worker has been diagnosed with an adverse health status (for example hypertension) and $Y(t)$ is an indicator that a worker has been diagnosed with the outcome of interest (for example ischemic heart disease). c is a time point subsequent to hire, at which point only active and disease-free workers are recruited into the cohort. All data sets contain K years of follow-up, but the prevalent cohort data, O_p, contains an additional c years of data prior to follow-up start.

The time-varying nodes are divided into two groups, where $A(t) = (E(t), C(t))$ are the intervention nodes and $L(t) = (H(t), L(t))$ are the non-intervention nodes. These are so named because we are interested in the effects of treatment regimes that possibly assign values to the intervention nodes, while allowing the non-intervention nodes to take the values that would result from the natural progression of the process under study.

We represent data histories using the overbar notation, so $\bar{L}(k) = (L(1), \ldots L(k))$ is the

history of L through time point k, and $\bar{A}(k) = (A(1), \ldots, A(k))$ is the history of A through k. We use $Q_{L(k)}$ to denote the conditional distribution of $L(k)$, given its parents. The parents of any variable X are all the variables that directly affect X; the parents of $L(k)$ are $Pa(L(k)) = (W, S, \bar{L}(k-1), \bar{A}(k))$. We use $g_{A(k)}$ to denote the conditional distribution of $A(k)$ given its parents $Pa(A(k)) = (W, \bar{L}(k-1), \bar{A}(k-1))$. We will also use the notation $g_{1:K} \equiv \prod_{j=0}^{K} g_{A(j)}$ and note that $g_{A(t)}$ can be factorized as $g_{E(k)}(E(k)|Pa(A(k)))g_{C(k)}(C(k)|Pa(A(k)), E(k))$.

2.2.2 Causal Model and Treatment Regimens

A causal model is a construct that serves as the link between the observed data and the counterfactual data that would have been observed if workers were exposed to a different treatment regimen. We use non-parametric structural equation models (Pearl, 2000) to construct our causal model. So, let

$$
\begin{aligned}
W &= f_W(U_W) \\
S &= f_S(U_S) \\
L(k) &= f_{L(k)}(Pa(L(k)), U_{L(k)}), k = 1 \ldots, c+K \\
A(k) &= f_{A(k)}(Pa(A(k)), U_{A(k)}), k = 1, \ldots c+K
\end{aligned}
$$

where $f_W, f_S, (f_{L(k)} : k = 1, \ldots c+K), (f_{A(k)} : k = 1, \ldots, c+K)$ are unspecified deterministic functions. $(U_W, U_S, U_{L(1)}, \ldots, U_{L(c+K)}, U_{A(1)}, \ldots, U_{A(c+K)})$ are sets of unmeasured background factors used by the functions to determine the data.

The causal relationships between these variables are pictured in the directed acyclic graphs (DAGs) (Pearl, 1995, 2000) in figures 2.1 and 2.2. The DAG in figure 2.1 illustrates the relationships between the baseline variables S and W and the intervention and non-intervention nodes $A(t)$ and $L(t)$. We note that S does not serve as an independent cause of the intervention nodes $A(t)$. The DAG in figure 2.2 illustrates the individual relationships between the time-varying covariates at two successive time points. This illustrates the role that $H(t)$ plays in the causal structure of our data, as both a confounder of the $E(t+1) \rightarrow Y(t+1)$ relationship and a mediator of the effect of $E(t)$ on $Y(t)$.

A post-intervention distribution is defined as the distribution that the observed data would have under a specified interventions that sets the values of $A(t)$ for $t = 1, \ldots, K$. The causal model is a model on all possible post-intervention distributions. The target causal parameters of interest are parameters of these post-intervention distributions, i.e. we are interested in making inferences about parameters of the distribution of non intervention variables under one or more possible interventions to set the intervention nodes. In general, we are interested in a contrast comparing the expectation of Y among the populations

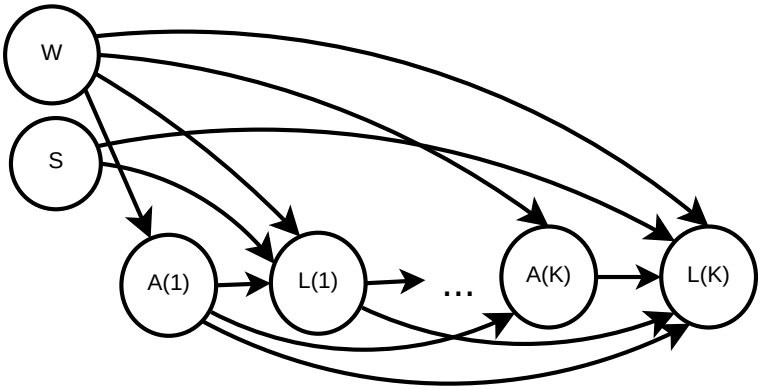

Figure 2.1: Directed acyclic graph illustrating the relationships between unmeasured susceptibility (S) measured baseline covariates (W) and the time-varying intervention ($A(t)$) and non-intervention ($L(t)$) nodes in the full data X.

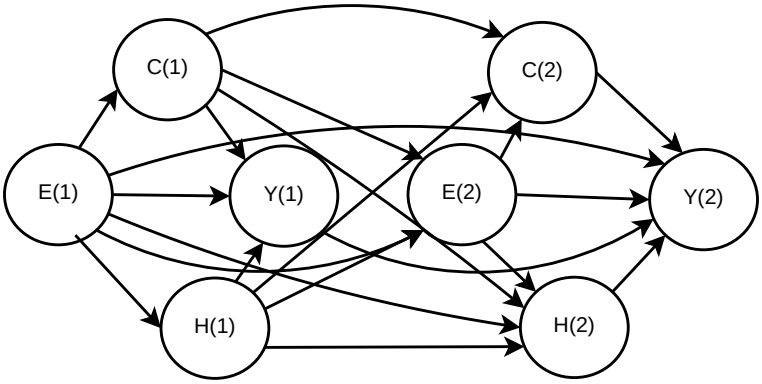

Figure 2.2: Directed acyclic graph illustrating the relationships between the time-varying covariates at two successive time points. The nodes represent binary exposure ($E(t)$), active work status ($C(t)$), diagnosis with an adverse health status ($H(t)$), and diagnosis with the outcome of interest ($Y(t)$).

following two complementary intervention regimes. For instance, we may be interested in the additive or the relative effects of exposure.

There are several types of target parameters that we consider in the course of this study, and each one defined with respect to a set of intervention regimes. One set are the static regimes $\{\bar{a}_0, \bar{a}_1\}$, which set binary exposure $E(t)$ to either 1 or 0 and $C(t)$ to 1 for all time points. We define the distribution of the non-intervention nodes under these regimes as $P_{\bar{a}_1}$ and $P_{\bar{a}_0}$. $P_{\bar{a}_1}$ and $P_{\bar{a}_0}$ are the distributions of the counterfactual data that would have been observed if all workers had been continuously exposed at these levels and stayed at work for the duration of follow up. The target parameters about which we would make inference are the means of the outcomes at time K, $E(Y_{\bar{a}_1}(K))$ and $E(Y_{\bar{a}_0}(K))$, parameters of the distributions $P_{\bar{a}_1}$ and $P_{\bar{a}_0}$.

Another set of static regimes that we consider are $\{\bar{e}_0, \bar{e}_1\}$, which set binary exposure $E(t)$ to either 1 or 0 for all point t, but do not intervene on leaving work $C(t)$. We define the distributions of the data under one of these regimes as $P_{\bar{e}}$. $P_{\bar{e}}$ is the distribution of the counterfactual data that would have been observed if exposures were controlled at a certain level in the workplace, but workers were allowed to leave freely. We would then be interested in making inference about the means at time K, $E(Y_{\bar{e}_1})$ and $E(Y_{\bar{e}_0}(K))$, parameters of the distributions $P_{\bar{e}_1}$ and $P_{\bar{e}_0}$.

Finally, we consider a class of dynamic regimes where a function $d(\bar{L}(t))$ of the observed data is used to set $A(t)$. We define d_1 as an intervention that sets $E(t)$ to 1 and $C(t)$ to 1 while the value of the time-varying health status $H(t-1) = 0$. However, once the value of $H(t-1)$ changes to 1, $d_1(\bar{L}(k))$ then sets $C(t)$ to 0 and therefore $E(t)$ to 0. d_0 is similarly defined except that active workers exposures are set to 0. We define the distributions of the observed data under these interventions as $P_{\bar{d}}$ and data following this distribution would be observed if exposure had been controlled at a certain level in the workplace, but sick workers were forced to leave work. Our corresponding parameters of interest for these types of regimes are $E(Y_{\bar{d}_1}(5))$ and $E(Y_{\bar{d}_0}(5))$. In all cases, regimes must be chosen by investigators both to generate causal contrasts of scientific interest and to ensure that workers have a positive probability of following the regimes of interest (see for example review in Petersen et al. (2012)).

2.2.3 Bias and Identifiability

The focus of this paper is to examine under what circumstances we will be able to generate unbiased estimates of these causal parameters. Under a set of assumptions, the statistical estimands generated from these observed data are equal to the target causal parameters. These assumptions are:

Positivity For any intervention $\bar{a}(t) = (a(1), \ldots, a(K))$, there is a positive probability that all workers could follow this intervention: $P(A(t) = a(t)|W, \bar{L}(t-1)) > 0 \ \forall W, t = 1, \ldots, K$.

Sequential Randomization Assumption The counterfactual values that the non-intervention nodes would take under each intervention are independent from the observed intervention nodes, given the observed covariates. $A(t) \perp L_d(t')|Pa(A(t)) \ \forall t, t' > t$ and regimes $d \in \mathcal{D}$. Here $\mathcal{D}$ is the set of all regimes we are interested in. This corresponds to there being no unmeasured confounders of any of the intervention nodes and non-intervention nodes of interest.

We use a conception of bias that encompasses both statistical bias as well as a divergence between the statistical parameter and the target parameter of interest caused by a lack of identifiability. One useful property of DAGs is their ability to allow researchers to determine whether effects of interest are identifiable, for example via the back-door criteria (Pearl, 2000) and its sequential analogue. For example, consider the DAG in figure 2.1, which illustrates the process that creates the incident cohort. The effect of $A(t), t = 1, \ldots, K$ on $L(K)$ is identifiable, as there are no unblocked paths between and $A(t)$ node and $L(K)$.

Next, consider the DAG in figure 2.3 which illustrates the selection into a prevalent cohort, where follow-up starts at time c. Here we split A into its component pieces, exposure E and active work status C and we are interested in the identifiability of the effect of both E and C on $L(K)$. Selection into the cohort involves conditioning on both $C(c-1) = 1$ and $Y(c-1) = 0$, which opens up a back-door path through $E(c) \leftarrow E(c-1) - - - S \rightarrow Y$. However, since $E(c-1)$ has been measured for each observed subject, this path is blocked and the effect is still identifiable. We therefore use the terminology left-truncation, as opposed to selection bias, to describe the creation of the prevalent cohorts, as the unblocked back-door paths opened by the selection process do not prevent the identifiability of the effect of interest.

An investigator might estimate the effect of a chosen intervention (for instance, $E(Y_{\bar{a}_1}(K))$, thecumulative disease incidence if all workers were exposed and remained at work for K years) among a prevalent cohort if this was the only cohort available for study. A natural interpretation of such an effect estimate would be that it is the effect of K years of exposure, without the provisio that it is only applicable to a prevalent cohort. However, this interpretation of taking the effect estimate and applying it to an incident cohort, is akin to attempting to identify the effect of a hypothetical intervention in which an incident cohort is turned into a prevalent cohort, through the prevention of disease and censoring.

Figure 2.4 contains a DAG in which we can examine the effect of such an intervention which sets $Y(t) = 0$ and $C(t) = 1$ for $t = 1, \ldots, c-1$ and then sets the intervention nodes for the K years following follow-up start $A(t), t = c, \ldots c + K$. That is, we would be interested in the additional effect of setting $C(t) = 1$ and $Y(t) = 0$ for $t = 1, \ldots, c$. In this case, the

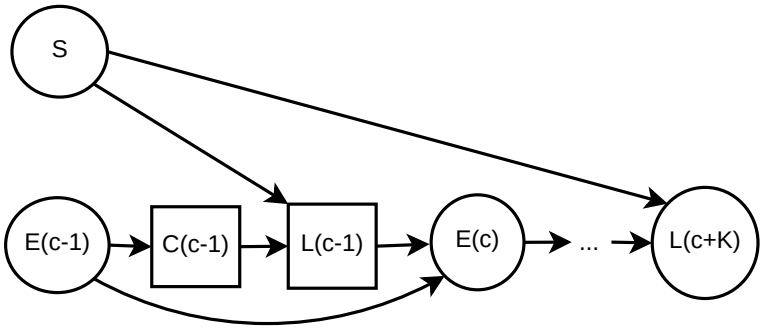

Figure 2.3: Directed acyclic graph illustrating the potential selection bias induced by conditioning on cohort membership. The prevalent cohort starts follow up at time point c with workers having $Y(c-1) = 0$ and $C(c-1) = 1$. This conditioning opens up a potential back-door path from $E(c-1)$ to S, which is however, blocked due to $E(c-1)$ having been measured.

intervention nodes include $L(c-1)$ as well as $A(c-1)$ and $A(t), t = c, \ldots c+K$. However, there is now an unblocked path through $L(c-1) \leftarrow S \rightarrow L(c+K)$, and the effect of this intervention is not identifiable. There are now unmeasured confounders between the intervention and non-intervention nodes.

The actual target parameter being estimated in the prevalent cohort is $E(Y_{\bar{a}_1}(c+K)|Y(c-1) = 0, C(c-1) = 1)$, while the corresponding parameter from an incident cohort is $E(Y_{\bar{a}_1}(K))$. Calculating an effect estimate in a prevalent cohort implicitly conditions on survival and remaining at work until follow-up starts, and this conditional density may be unidentifiable. Therefore, estimators summarizing the effect of exposure in an incident cohort using an observed prevalent cohort may get the wrong answer. This is not due necessarily due to any statistical property of the estimation procedure, but rather due to this lack of identifiability. To simplify our language, and in keeping with a common epidemiological practice, we will refer to differences in effect estimates caused by this lack of identifiability as bias.

2.2.4 Simulation

We designed a simulation study to explore the behavior of occupational cohorts and the performance of estimation procedures when applied to them. The simulated data sets contained the variables $X = (W, S, \bar{L}(t), \bar{A}(t))$, as defined earlier. Our simulated data sets were generated in accordance with the DAGs and using the following formulas:

Baseline Covariates $(W = (W_1, W_2),\ S)$ where $W_1 \sim \text{Bern}(0.75)$, $W_1 \sim \text{Bern}(0.7)$ and

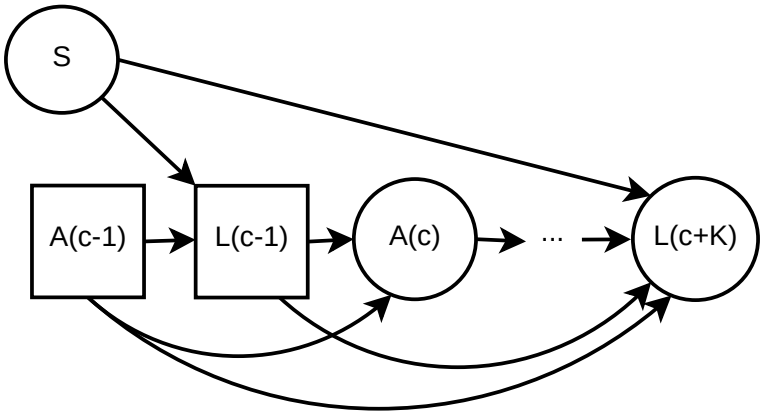

Figure 2.4: Directed acyclic graph illustrating the lack of identifiability of an intervention on nodes $A(c-1)$, $L(c-1)$ and $A(c)$. S is an unmeasured confounder between the intervention node $L(c-1)$ and outcome $L(K)$

$S \sim \text{Bern}(0.5)$ are independent from each other.

Exposure (E) where $E \sim \text{Bern}(\text{logit}([\beta_0^E + \beta_1^E W]I(t=1) + [\beta_2^E + \beta_3^E E(t-1) + \beta_3^E H(t-1)]I(t>1)))$, $\beta_0^E = -1.2, \beta_1^E = (1, 0.8), \beta_2^E = -0.3, \beta_3^E = 1$, and $\beta_4^E = -0.5$.

At Work (C) where $C(t) = 0$ if $C(t-1) = 0$. Otherwise, $C \sim \text{Bern}(\text{logit}(\beta_0^C + \beta_1^C W + \beta_2^C H(t-1) + \beta_3^C E(t)))$, $\beta_0^C = 3.5, \beta_1^C = (-0.1, -0.1), \beta_2^C = -0.5$, and $\beta_3^C = -0.5$.

Adverse Health Status (H) where $H(t) = 1$ if $H(t-1) = 1$. Otherwise, $H \sim \text{Bern}(\text{logit}(\beta_0^H + \beta_1^H W + \beta_2^H S + \beta_3^H S * E(t) + \beta_4^H E(t)))$, $\beta_0^H = -3, \beta_1^H = (0.5, -0.5), \beta_2^H = 0.5, \beta_3^H = 0.5$, and $\beta_4^H = 0.5$.

Outcome (Y) where $Y(t) = 1$ if $Y(t-1) = 1$. Otherwise, $Y \sim \text{Bern}(\text{logit}(\beta_0^Y + \beta_1^Y W + \beta_2^Y S + \beta_3^Y S * E(t) + \beta_4^Y H(t) + \beta_5^Y C(t) + \beta_6^Y E(t) + \beta_7^Y \sum_{k=1}^{t-1} E(k)))$, $\beta_0^Y = -4.5, \beta_1^Y = (-0.2, -0.2), \beta_2^Y = 0.5, \beta_3^Y = 0.5, \beta_4^Y = 0.75, \beta_5^Y = 0.5, \beta_6^Y = 0.1$, and $\beta_7^Y = 0.1$.

Data sets O_i were created by removing S from X. Data sets O_p were created by simulating cohort experience for $c+K$ years and including the histories for only workers with $C(c) = 1$ and $Y(c) = 0$. We designed three sets of experiments using these simulated data sets to illustrate different aspects of occupational cohort studies. All simulation and data analysis were performed in R 3.0.2 (R Development Core Team, 2008).

Our first set of simulation experiments explored the performance of standard methods of confounding adjustment compared to causal methods when time-varying confounding was active, but left-truncation was not. The observed data for these experiments consisted of $n = 10,000$ *iid* copies of O_i, and follow-up ended at employment termination. We first ran

a standard Cox regression analysis, estimating the parameters of the model:

$$\lambda(t|E(t), W) = \lambda_0(t)(\exp\{\alpha_1 E(t) + \alpha_2 W\}),$$

where $\lambda(t|E(t), W)$ and λ_0 are the conditional and baseline hazards, respectively.

Using the same data set, we next fit an inverse probability of treatment and censoring weighted (IPT/CW) Cox model (Hernán et al., 2000). This model is:

$$\lambda_{T_{\bar{a}_1}}(t|E(t), W) = \lambda_0 \exp\{\beta_1 E(t) + \beta_2 W\},$$

where $\lambda_{T_{\bar{a}_1}}$ is the hazard at time t among subjects with covariates values $E(t), W$ had they followed treatment regimen $\bar{A} = \bar{a}_1$, i.e. set $E(t) = 1$ and $Ct) = 0$ for $t = 1, \ldots, K$. Each worker-year was weighted by a stabilized weight

$$w_t = \frac{g_{1:t,n}(A(t)|W, \bar{A}(t-1) = \bar{a}(t-1))}{g_{A,1:t,n}(A(t)|W, \bar{L}(t), \bar{A}(t-1) = \bar{a}(t-1)))}.$$

where $g_{1:t,n}$ is the product of two independent logistic regressions fitting the exposure and censoring mechanisms, given the conditioning covariates. We evaluated the bias of the standard Cox model by comparing the model parameter estimate associated with exposure ($\hat{\alpha}_1$) to the value of the parameter returned by the IPT/CW model ($\hat{\beta}_1$), which we know to be unbiased in data structures such as these (Westreich et al., 2012).

We also analyzed the data using the longitudinal TMLE of a mean outcome procedure (LTMLE), which was introduced by van der Laan and Gruber in 2012 (van der Laan and Gruber, 2012). This procedure produces a doubly-robust loss-based substitution estimator of the cumulative incidence of Y at time point t if workers follow a specific exposure regimen. For the first set of experiments, we estimated $E(Y_{\bar{a}_1}(5))$ and $E(Y_{\bar{a}_0}(5))$, the cumulative incidences of Y if all workers are always exposed ($\bar{a}_1$) or unexposed ($\bar{a}_0$) while in both cases preventing censoring. We compared the estimates to the actual cumulative incidences obtained when simulating the data under each intervention regimen of interest. We used the `ltmle` package (Schwab et al., 2013, Lendle et al., 2014) to perform this analysis. We chose to focus on cumulative incidence estimators as generated by the LTMLE procedure as our metric of interest for the subsequent simulation studies.

For the second set of simulation experiments, we investigated estimator performance in the presence of left-truncation and susceptibility heterogeneity. The observed data used in this case was $n - m$ *iid* copies of O_p, where m was the total number of workers who left work or acquired the disease prior to time point 6, i.e. had either $C(t) = 0$ or $Y(t) = 1$ for $t = 1, \ldots, 5$. We used LTMLE to evaluate $E(Y_{\bar{a}_1}(10))$ and $E(Y_{\bar{a}_0}(10))$, the results of treatment regimens that set exposure and prevent leaving work, i.e. either $E(t) = 1$ or $E(t) = 0$ while $C(t) = 1$ for $t = (6, \ldots 10)$. Follow-up for these data ended at employment termination. These results were compared to the true cumulative incidences as determined

via simulation among an incident cohort, where subjects were followed up for the first five years after hire. We then refer to the difference between the estimate from the observed data and this truth from the simulation as the bias. This bias is primarily a function of differences between the surviving population that comprise a prevalent cohort and the original, incident, population from which they were drawn.

For the third set of simulation experiments, we allowed follow-up to continue after subjects had terminated employment. The observed data consisted of $n = 10,000$ iid copies of O_i, where follow-up did not end at employment termination. We then calculated estimators for $E(Y_{\bar{e}_1}(5))$, $E(Y_{\bar{e}_0}(5))$, $E(Y_{\bar{d}_1}(5))$, and $E(Y_{\bar{d}_0}(5))$ using the LTMLE procedure. $Y_{\bar{e}_1}$ and $Y_{\bar{e}_0}$ correspond to interventions that set exposure values while at work, but do not intervene on leaving work, while $Y_{\bar{d}_1}$ and $Y_{\bar{d}_0}$ correspond to interventions that enforce employees leaving work subsequent to diagnosis with health status $H = 1$. We then evaluated the performance of the LTMLE estimation procedure by comparing it to the true cumulative incidences as determined through simulation. These regimes correspond to setting exposure to 1 or 0 respectively unless a worker acquires the adverse health status, in which case she leaves work. For the simulations focusing on the intervention regimens d_0 and d_1, we set the values of the simulation parameters to $\beta_2^C = -2$ and $\beta_0^H = -3.5$ to ensure that the observed data contained sufficient workers following the defined intervention regimens.

2.3 Results

The results from the first set of experiments are contained in table 2.1. The estimands from the LTMLE procedure were generally unbiased. In contrast the estimands from the standard Cox model demonstrated a downwards bias. This bias makes exposure appear more protective from disease, whether the true exposure-response relationship was positive, null, or negative.

The second set of experiments explored the combined effect of left truncation and susceptibility heterogeneity and consisted of eight scenarios, described in table 2.2, with the results contained in table 2.3. Each scenario compares the estimands for $E(Y_{\bar{a}_1}(5))$ and $E(Y_{\bar{a}_0}(5))$ as generated by ltmle from the prevalent cohort with the truth as simulated from an incident cohort; with the differences reported as bias on both the relative and additive scales. The first two scenarios demonstrate the effects of left truncation when susceptibility is homogeneous versus heterogeneous in the population. We see no bias on the additive scale when susceptibility is homogenous (experiment 1), but we do see some when 50% of the workers are susceptible. We observe bias on the relative scale in both cases due to an overall higher risk in the left-truncated data. In scenario 3, when the levels of susceptibility are stronger, we see the bias concurrently increase.

Scenarios 4 and 5 investigate the effect of left truncation plus heterogeneous susceptibility

Table 2.1: Estimation results from the simulation study comparing estimator performance in a data set where a time-varying confounder is affected by prior exposure. The positive effect size was established by setting values of 0.5, 0.1, and 0.1 for β_4^Y, β_6^Y, and β_7^Y, the parameters corresponding to H, E and ΣE in the equation determining Y. The null exposure effect was established by setting each of these to 0, and the negative exposure effect was established by setting them to -0.5, -0.1, and -0.1, respectively. Reported biases are on the log scale, so a bias of -0.05 indicates that the log of the estimated rate ratio is -0.05 lower than the true value.

	Exposure	**LTMLE**			**Cox Model**		
Experiment	**Effect**	**True RR**	**Estimated RR**	**Bias**	**IPT/CW Estimate**	**Standard Estimate**	**Bias**
1	Positive	1.42	1.44	0.01	1.15	1.07	-0.07
2	Null	1.00	1.01	0.01	1.00	0.94	-0.05
3	Negative	0.71	0.73	-0.01	0.88	0.83	-0.05

Table 2.2: Description of simulation conditions. Weak susceptibility corresponds to values of 0.5 for β_2^H, β_3^H,β_2^Y and β_3^Y, which correspond to S or $S * E$ in the equations that determined Y and H. Strong susceptibility corresponds to values of 1.5. A positive exposure effect corresponded to values of 0.3, 0.3, and 0.03 for β_4^Y, β_6^Y, and β_7^Y, the parameters corresponding to H, E and ΣE in the equation determining Y. The null effect corresponded to values of 0 for all three, while the negative exposure effect had values of -0.3, -0.3 and -0.03, respectively

Scenario	**Description**	**Susceptibility Heterogeneity**	**Exposure Susceptibility**	**Disease Susceptibility**	**Susc. %**	**Exposure Effect**
1	Susceptibility Heterogeneity Off	FALSE	N/A	N/A	0	POS
2	Susceptibility Heterogeneity On	TRUE	WEAK	WEAK	50	POS
3	Strengthen suscepetbility effect	TRUE	STRONG	STRONG	50	POS
4	Null Effect	TRUE	WEAK	WEAK	50	NULL
5	Negative Effect	TRUE	WEAK	WEAK	50	NEG
6	Strong Effect in Subset	TRUE	STRONG	STRONG	10	POS
7	Subset: Disease Susceptibility	TRUE	NONE	STRONG	10	POS
8	Subset: Exposure Susceptibility	TRUE	STRONG	NONE	10	POS

when the exposure has a null and negative (protective) effect on the outcome. We observe similar results to experiment 1, in that the bias from the combination of left truncation and susceptibility heterogeneity is downwards, regardless of the direction of the exposure-response relationship. In scenarios 6,7 and 8, we reduced the proportion of susceptible workers to 10 % of the population. We also varied whether susceptibility functioned as disease susceptibility (increased risk regardless of exposure), exposure susceptibility (an effect modifier of exposure-disease) or both. We observed a downwards bias in all three scenarios. The bias was largest when both disease and exposure susceptibility were active, while the bias due to disease susceptibility was larger than that due to exposure susceptibility.

The third set of experiments explored the ability of the LTMLE estimation procedure to generate unbiased estimates when follow-up continued past employment termination. The results in table 2.4 show the estimands and true values as determined through simulation in an incident cohort for parameters corresponding to interventions $\{e_0, e_1\}$ (set exposure but not active status) and $\{d_0, d_1\}$ (set exposure and work status dependent on health status H). We see that the LTMLE procedure was able to generate unbiased estimates of all four target parameters as well as additive and relative contrasts of them.

2.4 Discussion

We describe the HWSE as composed of at least two distinct phenomena. The first, and widely recognized, is time-varying confounding affected by prior exposure. We demonstrate, using a standard analytic procedure, both the direction and size of the bias that results from the presence of time-varying confounding on the causal pathway (as suggested by (Steenland, 2013)). We observed a downwards bias under all possible true exposure-response levels: positive, null and negative, provided that the health status associated with the outcome of interest reduces the probability of future exposure. This bias can be corrected for using a variety of 'causal' methods.

The less widely recognized phenomenon, and the focus of our second set of simulation experiments, is left truncation in the presence of heterogeneity of susceptibility. The findings from our simulation study indicate that this phenomena results in effect estimates in a prevalent cohort lower than the true effect as measured in an incident cohort. It occurs because susceptible workers preferentially leave work or acquire the disease and the observed effect is therefore measured in a population relatively more resistant to the exposure and disease of interest. Observation of a worker occurs conditional on their surviving at work until follow-up start, and the distribution of the conditioning event depends upon an unmeasured variable S, and hence the original distribution is unidentifiable. Estimates derived from a cohort consisting of both incident and prevalent workers will naturally fall in between the effects within the incident and prevalent cohorts.

Table 2.3: Results of the experiments varying susceptibility properties. The true effect of exposure as determined by simulation of the experience of the incident cohort subject to either constant exposure or lack thereof, while preventing censoring. Longitudinal targeted minimum-loss-based estimation was used to estimate the cumulative incidence if subjected to the defined regimens as in the simulated prevalent observed data sets. Bias is defined as the difference between the truth from the incident cohort and the estimates from the prevalent cohort. The absolute differences as well as their percentages of the true effect size are reported.

	Incident: Truth		**Prevalent: Estimated**		**Rate Ratio**		**Additive Effect**	
Scenario	**Exposed**	**Unexposed**	**Exposed**	**Unexposed**	**Bias**	**% Bias**	**Bias**	**% Bias**
1	0.13	0.10	0.11	0.08	-0.08	- 20%	0.00	2%
2	0.28	0.15	0.26	0.11	-0.21	- 26%	-0.02	-11%
3	0.34	0.19	0.48	0.18	-0.34	-35%	-0.14	-48%
4	0.21	0.23	0.20	0.20	-0.10	-2200%	-.02	-2500%
5	0.14	0.22	0.14	0.20	-0.07	-21%	-0.02	-27%
6	0.17	0.11	0.19	0.10	-0.27	-41%	-0.04	-42%
7	0.16	0.12	0.14	0.10	-0.10	-28%	-0.01	-16%
8	0.16	0.10	0.14	0.14	-0.16	-24%	-0.00	-5%

Table 2.4: Results from experiments extending follow up past employment termination. Intervention regiments $\bar{e}$ correspond to setting exposure values while at work, but allowing the leaving work process untouched. Intervention regimens $\bar{d}$ correspond to setting exposure values while at work, but intervening to enforce leaving work after diagnosis with time-varying health status H

	Estimator		Truth		Rate Ratio		Additive Effect	
Intervention	Exposed	Unexposed	Exposed	Unexposed	**Bias**	**% Bias**	**Bias**	**% Bias**
$\bar{e}$	0.06	0.05	0.06	0.05	0.00	0%	0.00	0%
$\bar{d}$	0.03	0.02	0.03	0.02	0.03	6%	-0.00	-1%

We observed a negative bias from this phenomenon across a variety of parameter levels and simulation conditions. The magnitude of the bias increased with the strength of susceptibility and persisted when only a small proportion of the population was designated as susceptible. The bias occurred whether the true effect of exposure was positive, null or negative, and whether disease susceptibility, exposure susceptibility, or both were functioning. Left truncation in the presence of heterogeneity in unmeasured susceptibility can be considered a form of HWSE, because it results in observed associations between exposure, survival, health, and cohort membership that obscure the etiologic effect.

We can view this source of HWSE bias as an instructive example of the concept of transportability. The effect that is estimated in the left-truncated population is not transportable to the original population, due to differences in their baseline proportions of susceptibility. While this concept is often applied to populations that differ in location, it is equally valid to apply it to incident and prevalent cohorts that are distinguished only by time.

In several papers, Bareinboim and Pearl (Pearl and Bareinboim, 2011, Bareinboim and Pearl, 2012) have given transportability a formal definition and demonstrated the use of DAGs to identify systems whose measured effects are transportable to each other, using the property of 'S-admissibility'. S-admissibility can be identified from a DAG by: (1) removing all of the arrows out of the exposure and (2) checking for unblocked pathways between the S node and the outcome (Petersen, 2011). If there are no pathways blocked by measured variables, than this set of measured variables is S-admissible and effects measured in one are transportable between populations. In their work, the S node stands for selection and represents the variables that differ between the populations. In our case, S conveniently stands for susceptibility, and a cursory look at Figure 2.1 demonstrates that S-admissibility does not hold for systems such as we describe, where S is a direct cause of the outcome.

Other discussions of transportability (Hernán and VanderWeele, 2011) have identified the fact that effect modifiers must be similarly distributed among the two populations in order for effect estimates to be transportable between them. The results of our simulation, and the use of the S-admissability criterion, show that this restriction is not solely applicable to effect modifiers of the exposure, but to direct causes of the outcome as well. We saw bias with respect to the incident cohort when susceptibility functioned as exposure susceptibility or disease susceptibility, though the bias was highest when it functioned as both. This analysis confirmed the findings of Applebaum et al. (2011), and placed them in the context of causal inference by using the concepts of identifiability and transportability to explain the observed distinctions between effect estimates.

Terminated employees cannot be exposed, so they represent a potential violation of the assumptions of positivity: a combination of covariates such that the probability of being exposed is 0. This violation implies the non-identifiability of causal effects defined by continuous treatment, which the parameters of a simple marginal structural model relating exposure and outcome correspond to. We chose to implement longitudinal TMLE of a mean

outcome in order to control for time-varying confounding on the causal pathway. The use of this estimation procedure also demonstrates the potential for estimation using a dynamic treatment regimen when follow-up extends past employment termination.

Picciotto et al. (2013) address the effect of truncating follow-up at employment termination. They concluded that it represented a potential source of bias in an applied data example, due to a lack of exchangability between the terminated and non-terminated workers. In this simulation, we did not observe the same bias because we had a measurement of the time-varying health status (H) that predicted employment termination, and thus the terminated and non-terminated workers were conditionally exchangeable.

Positivity violations have been offered (Robins, 2000b, Naimi et al., 2014) as a justification for the exclusive use of G-estimation of the parameters of a structural nested model when investigating occupational exposures. The parameters of a structural nested model do not rely on the assumption of positivity to be identifiable, as the model itself specifies the relationship between the timing of exposure and the timing of the outcome. In the simplest case of an accelerated failure time model, the model implies that the effect of exposure on the outcome is the same no matter when in an individual's employment history it occurs. While this assumption may be reasonable in some cases, in other cases it may not, so the viability of other estimation approaches is worth exploring.

By defining treatment regimens, whether static like $\bar{e}$ or dynamic like $\bar{d}$, we were able to generate estimates of the causal effect of exposure even when follow up continued past employment termination. We note, however, that a portion of the effect of exposure $E(t)$ on the outcome travels through the nodes $C(t)$ and $E(t+1)$. Indeed these effects of exposure (to reduce future exposure and induce employment termination) are the root of the HWSE that we are trying to remove from the analysis. We must be mindful, however, that the effect we estimate is not the total effect of E, as some of its effect moves through the pathways $E \rightarrow C$ or $E \rightarrow H \rightarrow E$. It is important to identify intervention regimes that correspond to scientific questions of interest but also to recognize that an effect estimate's interpretation is dependent on how the defined regimen corresponds to a realistic sequence of events in the real world. We would order the interventions we consider as $(\bar{a}, \bar{e}, \bar{d})$ from least to most realistic, as it is within the ability of an employer to control exposure levels, but less so to control when employees choose to leave work.

Chapter 3

Efficient Estimation in Data Structures with Missing Confounders

3.1 Introduction

This chapter contains an exploration of the performance of an augmented inverse weighted targeted minimum-loss based estimator for use in data structures with missing confounders, where the probability of the confounders being measured is a function of the entire history of the cohort. This section is organized as follows. In section 3.1, we describe the data structures of interest, list our assumptions, and demonstrate the identifiability of our target parameters. Section 3.2 contains background on other approaches to estimation for this problem, focusing on the methodology we will use for comparison and those that directly precede this paper. Section 3.3 describes the full data TMLE, the IPCW-TMLE, and the A-IPCW TMLE procedures and highlights some of the statistical properties of the A-IPCW TMLE estimators. Section 3.4 contains a description of and the results from a simulation study, which investigated estimator performance as the proportion of missing data varies as well as when various nuisance parameters were estimated based on misspecified models. Section 3.5 demonstrates the application of these estimators to an applied example, an investigation of the effect of occupational exposure to airborne particulate matter on the incidence of heart disease in a cohort of smelter workers.

3.2 Model and Identifiability

Let $X = (L, V)$ be the full data and $O = (RL, V, R)$ be the observed data. V is a vector of always measured covariates and $(A, Y) \subseteq V$ denote the treatment (A) and outcome (Y) of

interest. L is a vector of potential confounders, possible causes of both A and Y, and R is an indicator that the vector L has been measured. We are interested in some summary measure of the effect of the treatment on the outcome. As an example, the target parameter of interest might be following estimand, which under additional causal assumptions corresponds to the average treatment effect (ATE):

$$\psi^F = E_{W,L}(E(Y|A=1,W,L) - E(Y|A=0,W,L)).$$

where $W \subset V$ are the set of measured baseline confounders. We will return to this example at times, but this work is developed in generality in order to highlight its applicability to a variety of target parameters.

The full data model is non-parametric: $X \sim P_X, P_X \in \mathcal{M}^F$, where $\mathcal{M}^F$ is non-parametric. We indicate the true distribution of X with $P_{X,0}$ and note that it can be factorized as $P_{X,0} = P_{L|V,0}P_{V,0}$. We use the subscript 0 to refer to the true values of the parameters of an object and the subscript n to refer to parameters that are estimates based on the observed data. The mechanism generating R may be known or unknown, but we make two assumptions about it:

Missing At Random (MAR): $P(R|X) = P(R|V)$

Positivity: $P(R=1|V) > 0$ *a.e.*

These assumptions allow the factorization of the observed data likelihood as follows. Let $\mathcal{C}_X(O_i)$ be the coarsening of the observed data point $O_i = (R_iL_i, V_i, R_i)$ (i.e. $\mathcal{C}_X(O_i) = (V_i, L_i)$ if $R_i = 1$, otherwise $\mathcal{C}_X(O_i) = \{(V_i, l) : P_{L|V}(l|V_i) > 0\}$). The likelihood of an observed data point O_i is then:

$$\begin{aligned}
P(O = O_i) &= \int_{x \in \mathcal{C}_X(O_i)} \int_{r=R_i} P_X(X=x)P(R=r|X=x)\partial\nu_R(r|x)\partial\nu_X(x) \\
&= \int_{x \in \mathcal{C}_X(O_i)} P_X(X=x)P(R=R_i|X=x)\partial\nu_X(x) \\
&= \int_{x \in \mathcal{C}_X(O_i)} P_X(L=l, V=V_i)P(R=R_i|V=V_i, L=l)\partial\nu_X(x) \\
&= P(R=R_i|V=V_i) \int_{x \in \mathcal{C}_X(O_i)} P_X(L=l, V=V_i)\partial\nu_X(x) \qquad (3.1)
\end{aligned}$$

Equation 3.1 implies that the log of the observed data likelihood equals the sum of two terms

$$\log(P(O)) = \log(P(R|V)) + \log\left(\int_{x \in \mathcal{C}_X(O)} P_X(x)\partial\nu_X(x)\right)$$

only one of which has to do with P_X, so maximization of the observed data likelihood with

respect to parameters of P_X involves only the full-data likelihood.

We use the notation $\Pi(V)$ for $P(R=1|V)$, which identifies the conditional distribution of R. We denote the observed data likelihood with P_{Π,P_X} because, as shown in equation 3.1, the choice of Π and P_X identify this distribution. We assume the existence of a mapping Ψ^F from the full data distribution to the d dimensional reals: $\Psi^F : P_X \to \mathbb{R}^d$. The target parameter to be estimated is this mapping applied to the true distribution, $\Psi^F(P_{X,0}) = \psi_0^F$, and we define $D^F(P_X)$ as the efficient influence curve of ψ^F at $P_X \in \mathcal{M}$. For our example target parameter of the ATE, the efficient influence curve is:

$$\begin{aligned} D^F(P_X)(X) = \quad & \left(\tfrac{I(A=1)}{g(1|W,L)} - \tfrac{I(A=0)}{g(0|W,L)}\right)(Y - \bar{Q}(W,L,A)) \\ & + \bar{Q}(W,L|A=1) - \bar{Q}(W,L|A=0) - \psi^F. \end{aligned} \tag{3.2}$$

where $\bar{Q}(W,L,A) = E(Y|W,L,A)$ and $g(a|W,L) = P(A=a|W,L)$. Our goal is to find a mapping Ψ from the observed data distribution such that $\Psi(P_{\Pi,P_X}) = \Psi^F(P_X)$, where P_X on the right hand side is the full data distribution implied by P_{Π,P_X}.

We now demonstrate the identifiability of such a mapping from the observed data model $O \sim P_{\Pi_0,P_{X,0}}$, where $P_{\Pi_0,P_{X,0}}$ is the true observed data distribution and $P_{\Pi_0,P_{X,0}} \in \mathcal{M} = \{P_{\Pi,P_X} : P_X \in \mathcal{M}^F, \Pi \text{ has MAR and positivity}\}$.

$$\begin{aligned} P(V=v, L=l, R=1) &= P(R=1|V=v, L=l)P(V=v, L=l) \\ &\overbrace{=}^{\text{MAR}} P(R=1|V=v)P_X(V=v, L=l) \\ P_X(V=v, L=l) &\overbrace{=}^{\text{Positivity}} \frac{P(R=1, L=l, V=v)}{\Pi(V=v)} \end{aligned} \tag{3.3}$$

Both $P(R=1, L=l, V=v)$ and $\Pi(V)$ are identified from the observed data distribution. We can therefore define $\tilde{P}_X$ as an identifiable distribution with density $\frac{P(R=1,L=l,V=v)}{\Pi(V=v)}$ at all x. Hence there exist mappings, Ψ, from the observed data distribution to the reals equivalent to our desired full data mapping

$$\Psi(P_{\Pi,P_X}) \equiv \Psi^F(\tilde{P}_X) = \Psi^F(P_X).$$

We observe n *i.i.d.* copies of the random variable O_i and are concerned with the construction of estimators of ψ_0^F based on $\bar{O} \equiv (O_1, \ldots O_n)$.

3.3 Background

The presence of missing covariates is a ubiquitous problem in epidemiologic, clinical and social research (Greenland and Finkle, 1995), and has long been an area of active statistical and applied research. Likelihood based estimation, multiple imputation, and inverse weighting are all possible approaches to estimation in this situation. Likelihood-based estimation proceeds from the log of the factorized likelihood:

$$\log(P(O)) = \log(P(R|V)) + \log\left(\int_{x \in C_X(O_i)} P_X(V_i, l)\partial\nu_X(x)\right),$$

and then maximizes the empirical mean of the second term over $P_X \in \mathcal{M}$ (Rubin, 1976b, Kenward and Molenberghs, 1998). So,

$$P_{X,n} = \underset{P_X \in \mathcal{M}}{\operatorname{argmax}} \sum_{i=1}^{n} \log\left(P_X(X_i)\right)^{R_i} + \log\left(\int_{x \in C_X(O_i)} P_X(V_i, l)\partial\nu_X(x)\right)^{1-R_i}, \tag{3.4}$$

and the maximum likelihood estimator is $\psi_n = \Psi(P_{X,n})$. This is an attractive approach because of the optimality properties of the MLE, (Le Cam et al., 1986) but it is rarely used in modern applications because if the dimension of L is high or contains continuous components, the MLE becomes ill defined for the non parametric model.

Multiple imputation is viewed as an accessible and widely applicable approach to estimation when confounders are missing (Rubin, 1996, Klebanoff and Cole, 2008). First proposed by Rubin (1987), it is a process in which $m = 1, \ldots M$ full data sets, $\bar{X}^{(m)} = (X_1^{(m)}, \ldots X_n^{(m)})$ are created. There are a variety of methods proposed for accomplishing this, but we describe only the 'proper' Bayesian approach (Nielsen, 2003). This begins with a parameterization of the conditional density of the missing covariates, $P_{L|V}$, by θ, i.e. $P_{L|V}(L|V;\theta)$, as well as assumption of an (often non-informative) prior distribution, $P(\theta)$. The conditional distribution of θ given the observed data, $P(\theta|\bar{O})$, is then estimated by combining $P(\theta)$ with the conditional distribution of the data $P(\bar{O}|\theta)$ using Bayes' rule. For each of the $m = 1, \ldots, M$ desired data sets, draws $\theta^{(m)}$ are made from $P(\theta|\bar{O})$ and $P_{L|V}(L|V;\theta^{(m)})$ is then used to generate $L^{(m)} = (L_i^{(m)})$ for all subjects i with $R_i = 0$. $L^{(m)}$ is combined with O to create a 'full' data set $\bar{X}^{(m)}$.

We assume the existence of an estimation algorithm, $\hat{\Psi}(\bar{X})$ that generates a regular and asymptotically linear estimator of $\Psi^F(P_X)$. This full data estimation algorithm $\hat{\Psi}^F()$ can then be applied to generate $\psi_n^{(m)} = \hat{\Psi}^F(\bar{X}^{(m)})$ and variance estimate $\sigma_n^{2,(m)}$. Rubin's rules (Schafer, 1997) can be used to combine these to generate estimates of the target parameter

and its variance. So,

$$\psi_{n,mi} = M^{-1} \sum_{m=1}^{M} \psi_n^{(m)}$$

and the corresponding variance estimate is

$$\sigma_n^2 = M^{-1} \sum_{m=1}^{M} \sigma_n^{2,(m)} + (1 + M^{-1})(M)^{-1} \sum_{m=1}^{M} (\psi_{n,mi} - \psi_n^{(m)})^2.$$

Multiple imputation can be seen as a type of maximum likelihood approach that avoids the need to compute difficult integrals such as in equation 3.4 (Carpenter et al., 2006, Seaman and White, 2011). The multiple imputation literature is rich and varied (S Su et al., 2011, L Schafer and W Graham, 2002, White et al., 2011), with different strategies proposed to deal with data structures with differing missingness patterns, variable types comprising L, and relationships $P_{L|V}(L|V;\theta)$. These methods commonly make parametric assumptions on the shape of $P_{L|V}$, and therefore assume a more restrictive model than $\mathcal{M}$.

Inverse-probability weighted based estimation is an alternative approach to this problem, whose form is suggested by the identifiability result in equation 3.3. Each data point is weighted by $\frac{I(R_i=1)}{\Pi_n(V_i)}$, where $\Pi_n(V)$ is an estimate of the probability that a subject would have been completely observed. The full data estimation procedure is then applied to this reweighted dataset. This process requires an estimation procedure that can accept a weighted data set and estimators of this type generally suffer from a lack of efficiency (Li et al., 2011, Seaman and White, 2011). This can be understood heuristically by noting that only complete case subjects contribute information on the relationship between A and Y to the estimation procedure, while subjects with $R_i = 0$ contribute no information beyond the fit of Π_n.

In a series of articles (Robins et al., 1994, Robins and Rotnitzky, 1995, Rotnitzky and Robins, 1995), Robins and colleagues introduced a class of inverse weighted estimators in a semi-parametric regression model, which was subsequently fully developed for any full data model (van der Laan and Robins, 2003). They define $D^F(\psi, h)$ as a class of full data estimating functions for ψ indexed by functions h. These can be mapped into observed data estimation functions $D(\psi, h, \phi)$ indexed by h and ϕ. They define a class of estimators as solutions to the corresponding estimating equations: $0 = \bar{D}(\psi, h, \phi)(\bar{O}) = n^{-1} \sum_i D(\psi, h, \phi)(O_i)$, where

$$D(\psi, h, \phi)(O_i) = \frac{R_i}{\Pi(V_i)} D^F(\psi, h) - \left(\frac{R_i}{\Pi(V_i)} - 1 \right) \phi,$$

where D^F is a gradient of ψ^F.

The efficient choice for ϕ is $\phi^h = E(D^F(\psi_0, h)|V, R = 1)$. We will use the notation, γ, for the conditional expectation of a full data gradient, and we note that, given that the MAR

assumption holds,

$$\gamma \equiv E(D^F(\psi, h)|V, R = 1) = E(D^F(\psi, h)|V)$$

Estimators in this class are called 'double-robust', because they remain consistent if either Π_n or γ_n are consistent. Estimators of this type have been found to be computationally challenging (Carpenter et al., 2006, Williamson et al., 2012) and their implementation in the applied literature has been limited, in part because the form of the expectation $E(D^F(\psi_0, h)|V)$ is generally unknown. Some authors have suggested forgoing the estimation of γ and instead relying on flexible modeling strategies such as splines (Little and Hyonggin, 2003) or tree-based algorithms (Li et al., 2011) to generate consistent estimators of Π_0.

Rose and van der Laan (Rose and van der Laan, 2012) considered the application of targeted minimum-loss based estimation (TMLE) within model $\mathcal{M}$. We fully describe this estimation process in section 4.2, which relies upon an estimate Π_n of the probability that a subject is fully observed. Rose and van der Laan recommend the nonparametric estimation of Π_n, but note that when this is not feasible, a targeting step can be applied to Π_n to ensure that efficient influence curve is solved. The focus of this paper is to implement and evaluate this approach; we update the fit of Π_n in order to create efficient and doubly-robust estimators of ψ_0. We will refer to this approach as augmented inverse probability of censoring weighted targeted minimum loss-based estimation (A-IPCW TMLE).

3.4 A-IPCW TMLE

3.4.1 Full-Data Targeted Minimum Loss Based Estimation (TMLE)

TMLE is a generalized method for the creation of loss-based efficient substitution estimators. It was first introduced by van der Laan and Rubin (2006a) and van der Laan and Rose (2011) present a comprehensive description and demonstration of the methodology within a large class of estimation problems. TMLE involves the minimization of the empirical mean of loss functions with respect to specifically defined parametric submodels, which generate estimates of components of the observed data distribution that are targeted to the parameter of interest. For the current work, we assume that a full-data TMLE for the parameter of interest has been defined. That is, given n *iid* observations $\bar{X} = (X_1, \ldots, X_n)$, there exists a mapping $\hat{\Psi}(\bar{X}) = Q^*_{X,n}$, an estimator for $Q_X(P_{X,0})$, which are components of the full data distribution $Q_X(P_X)$ such that $D^F(Q_X) = D^F(P_X)$ and $\Psi^F(Q_X) = \Psi^F(P_X)$. For example, if the estimand corresponding to the ATE is the target parameter, with D^F defined as in equation 3.2, the set of components Q_X so defined are $(P_{W,L}, g, \bar{Q}, \psi)$. Given a a targeted estimate of these components $Q^*_{X,n}$, $\Psi^F(Q^*_{X,n})$ is the corresponding TMLE of the target parameter ψ^F_0.

TMLE requires $L^F(Q_X)$, a full data loss function minimized by the true distribution:

$$Q_{X,0} = \underset{\{Q_X(P_X):P_X\in\mathcal{M}^F\}}{\operatorname{argmin}} E_0 L^F(Q_X)(X)$$

Let $\{Q_X(\epsilon) : \epsilon\}$ be a working parametric submodel of Q_X constructed so that its score at $\epsilon = 0$ equals the full-data efficient influence curve:

$$\frac{\partial}{\partial\epsilon}\log(Q_X(\epsilon)(X))\bigg|_{\epsilon=0} = D^F(Q_X)(X) \tag{3.5}$$

Starting with an initial estimate of the components of the full data density, $Q^0_{X,n}$, ϵ^k_n is defined as the iterative minimizer of the full-data loss function applied to the submodel of the component estimate $k-1$:

$$\epsilon^k_n = \underset{\epsilon}{\operatorname{argmin}}\, P^F_n L^F(Q^{k-1}_{X,n}(\epsilon)) \tag{3.6}$$

and then $P^k_{X,n} = Q^{k-1}_{X,n}(\epsilon^k_n)$. Above, P^F_n is the full-data empirical distribution of $\bar{X}$, which is not identifiable from the observed data when some covariates are subject to missingness, thus motivating our current study. Equation 3.6 is iteratively solved for $k = 1, \dots K$ until $\epsilon^K_n \approx 0$. Minimizing the loss function with respect to the submodel $Q_X(\epsilon)$ ensures that the empirical mean of equation 3.5, is zero. Therefore, the full-data TMLE solves the estimating equation associated with the full data efficient influence curve

$$P^F_n D^F(Q^*_{X,n}) = 0$$

and achieves the minimal variance among unbiased estimators of ψ_0 in $\mathcal{M}^F$.

3.4.2 IPCW-TMLE

Simple IPCW-TMLE, as introduced by Rose and van der Laan (2012), works by the incorporation of a weight: $\frac{R}{\Pi(V)}$ into the process described above. Specifically, the observed data loss function is defined as:

$$L(Q_X)(O) = \frac{R}{\Pi(V)}(L(Q_X)(X))$$

A parametric submodel $Q_X(\epsilon)$ is then defined so that

$$\frac{\partial}{\partial\epsilon}\log Q_X(\epsilon)\bigg|_{\epsilon=0} = \frac{R}{\Pi(V)}D^F(Q_X)(X) \tag{3.7}$$

and ϵ is estimated as the minimizer of the empirical mean of the loss function applied to that submodel

$$\begin{aligned}\epsilon_n^k &= \underset{\epsilon}{\operatorname{argmin}}\, P_n L(Q_{X,n}^{k-1}(\epsilon)) \\ &= \underset{\epsilon}{\operatorname{argmin}}\, n^{-1} \sum_i \frac{R_i}{\Pi(V_i)} L(Q_{X,n}^{k-1}(\epsilon))(X_i). \end{aligned} \tag{3.8}$$

The likelihood estimate is iteratively updated, with $Q_{X,n}^k \equiv Q_{X,n}^{k-1}(\epsilon_n^k)$ for $k = 1, \ldots, K$ until $\epsilon_n^K \approx 0$. The final estimator $Q_{X,n}^{K-1}(\epsilon_n^K) \equiv Q_{X,n}^*$ is a targeted estimate of $Q_{X,0}$ and $\Psi^F(Q_{X,n}^*) = \psi_n$ is the IPCW TMLE for ψ_0. This fit, $Q_{X,n}^*$ solves the efficient score equation for the full data parameter, weighted by the inverse probability of being fully observed.

$$P_n \left(\frac{R}{\Pi_n(V)} D^F(P_{X,n}^*) \right) = 0 \tag{3.9}$$

Since the empirical score of $Q_{X,n}^*$ with respect to ϵ is 0, and equation 3.9 holds, the IPCW-TMLE solves the estimating equation corresponding with the inverse weighted full data efficient influence curve.

Rose and van der Laan prove that any additional properties of the full data TMLE (such as double-robustness to the misspecification of the treatment and outcome mechanisms), are inherited by the IPCW-TMLE, as long as the estimate of the missingness mechanism, Π_n, is consistent. They also prove that if Π_n is estimated using non-parametric maximum likelihood, then the IPCW-TMLE is an efficient estimator of ψ_0. The reasons for this become clear upon examination of the efficient influence curve for ψ^F in model $\mathcal{M}$, which is:

$$D^*(P_X, \Pi)(O_i) = \frac{R_i}{\Pi(V_i)} D^F(P_X)(X_i) - \frac{R_i - \Pi(V_i)}{\Pi(V_i)} E(D^F(P_X)|V_i, R_i = 1). \tag{3.10}$$

If Π_n is the non-parametric maximum likelihood estimator, it solves the score equations for all parametric submodels of Π_n (Gill et al., 1989). The score equations for submodels

$$\operatorname{logit} \Pi(\delta) = \operatorname{logit} \Pi + \delta(\phi(V)),$$

univariate logistic regressions with covariate $\phi(V)$, are:

$$\bar{S}(\phi) = \sum_{i=1}^{n} (R_i - \Pi_n(V_i))\phi(V_i)$$

If Π_n is the NPMLE, then $\bar{S}(\phi) = 0$ for all covariates $\phi(V)$, in particular for $\phi(V) = \frac{E(D^F(P_X)|V,R=1)}{\Pi_n(V)}$. Therefore the empirical sum of the second component of equation 3.10 will be 0. Combined with the fact that the IPCW-TMLE procedure ensures that equation 3.9

holds, we have that

$$P_n D^*(P_{X,n}, \Pi_n)(X) = 0.$$

However, if non-parametric MLE is not feasible (i.e. if V is high-dimensional and/or has continuous components), additional steps must be taken to achieve efficient estimation. As outlined by Rose and Van der Laan (Rose and van der Laan, 2012), this can be achieved by the targeted estimation of the missingness mechanism, Π, incorporating an estimate of the nuisance parameter γ. We now describe an approach to this targeting step, the central component of the A-IPCW TMLE estimation process.

3.4.3 Augmented-IPCW TMLE

Augmented IPCW TMLE incorporates an estimator γ_n for $E(D^F(P_X)|V, R = 1)$ into the fitting process for the missingness mechanism, Π_n. From here on, we will incorporate γ into our notation for influence curves, i.e.

$$D(P_X, \Pi, \gamma) = \frac{R}{\Pi(V)} D^F(P_X) - \frac{R - \Pi(V)}{\Pi(V)} \gamma(V).$$

The augmentation procedure iteratively updates $\Pi_n(V)$ until arriving at a final augmented estimator for the missingness mechanism, Π_n^*. Π_n^* guarantees that the resulting estimates of likelihood components $Q_{X,n}^*$ solve $\sum_i D_i^*(Q_{X,n}^*, \Pi_n^*, \gamma_n) = 0$. This result implies that if γ_n is consistently estimated, the A-IPCW TMLE estimator, $\psi_n = \Psi(P_{X,n}^*)$, achieves the minimal variance of all unbiased estimators of ψ_0 in $\mathcal{M}$.

The iterative updating of Π_n starts with an initial estimate of the missingness mechanism, denoted $\Pi_n^0(V)$, and an estimate $\gamma_n(V)$ of the regression of $D^F(P_X)$ on V. Any procedure the investigator favored could be used to generate these estimates, for example a loss-based ensemble learner such as SuperLearner (van der Laan et al., 2007). For $k = 1, \ldots, K$, a parametric submodel $\Pi_n^k(\beta)$ is defined such that

$$\Pi_n^k(\beta)(V) = \text{expit}\left(\text{logit}(\Pi_n^k(V)) + \beta\left(\frac{\gamma_n(V)}{\Pi_n^k(V)}\right)\right) \tag{3.11}$$

The empirical mean of the logistic loss function $L(\Pi)(O) = -\log(\Pi(V))^R(1 - \Pi(V))^{1-R}$ is then minimized over $\bar{O}$ to estimate β_n^k.

$$\beta_n^k = \underset{\beta}{\text{argmin}}\, P_n L(\Pi_n^{k-1}(\beta)) \tag{3.12}$$

This corresponds to running a logistic regression of R_i on $\frac{\gamma_n(V_i)}{\Pi_n^{k-1}(V_i)}$, using $\Pi_n^{k-1}(V_i)$ as an offset. The subsequent fit is defined as $\Pi_n^k = \Pi_n^{k-1}(\beta_n^k)$ and the process is iteratively updated

for $k = 1, \dots K$, until convergence when $\beta_n^K \approx 0$. The final estimate of the missingness mechanism is denoted $\Pi_n^* \equiv \Pi_n^{K-1}(\beta_n^K)$. This estimate Π_n^* is then used in the IPCW-TMLE procedure to generate an estimate of the likelihood component $Q_{X,n}^*$. As described above, this involves weighting the full data TMLE loss function and submodel by $\frac{R}{\Pi_n^*(V)}$ and iteratively minimizing the submodel until convergence at the final estimator, $Q_{X,n}^*$. The A-IPCW TMLE is then $\psi_n = \Psi^F(Q_{X,n}^*)$.

A-IPCW solves the Efficient Influence Curve

The contribution to the observed data likelihood (equation 3.1) from $\Pi_n^k(\beta)$ is (allowing E_i to stand for $\frac{\gamma_n(V_i)}{\Pi_n^k(V_i)}$ for notational clarity):

$$\mathcal{L}(\Pi(\beta)) = \prod_{i=1}^{n} \frac{\exp(R_i(\beta E_i)}{1 + \exp(\beta E_i)}$$

with a corresponding score with respect to β of:

$$\begin{aligned}
\mathcal{S}(\Pi(\beta)) = \frac{\partial \log(\mathcal{L}(\Pi(\beta)))}{\partial \beta} &= \frac{\partial}{\partial \beta} \sum_{i=1}^{n} R_i \beta E_i - \log(1 + \exp(\beta E_i)) \\
&= \sum_{i=1}^{n} R_i E_i - \frac{\exp(\beta E_i)}{1 + \exp(\beta E_i)} E_i \\
&= \sum_{i=1}^{n} \frac{E(D^F | V_i, R_i = 1))}{\Pi_n^k(V_i)} (R_i - \Pi_n^k(\beta)(V)) \\
&\text{and when } \beta_n^K \approx 0 \text{ and therefore } \Pi_n^k(\beta) \approx \Pi_n^k \\
&= \sum_{i=1}^{n} \frac{R_i - \Pi_n^k(V_i)}{\Pi_n^k(V_i)} E(D^F | V_i, R_i = 1). \qquad (3.13)
\end{aligned}$$

The use of the logistic loss function ensures that the empirical score, equation 3.13, equals 0. Therefore, an updated Π_n will result in an estimator that also ensures that the empirical sum of the second half of the efficient influence curve is 0. When combined with the property of IPCW TMLE that guarantees that equation 3.9 holds, we have that

$$P_n D^*(Q_{X,n}^*, \Pi_n(\beta), \gamma_n) = 0. \qquad (3.14)$$

A-IPCW TMLE is Double Robust

The following section demonstrates that the A-IPCW estimator is doubly-robust to the misspecification of either Π_n or γ_n, as previously established in e.g. in van der Laan and Robins (2003). We first note that any influence function of the form $D^*(Q_X, \Pi, \gamma)$ can be re-expressed as follows:

$$\begin{aligned} D^*(Q_X, \Pi, \gamma) &= \left(\frac{R}{\Pi(V)} D^F(Q_X)(X) - \left(\frac{R}{\Pi(V)} - 1\right)\gamma(V)\right) \\ &= D^F(Q_X)(X) + \left(\frac{R-\Pi(V)}{\Pi(V)}\right)\left(D^F(Q_X)(X) - \gamma(V)\right). \end{aligned} \tag{3.15}$$

Therefore,

$$P_0 D^*(Q^*_{X,n}, \Pi_n, \gamma_n) = P_0 D^F(Q^*_{X,n}) + P_0 R(\Pi_n, \gamma_n(Q_X))$$

where $R(\Pi_n, \gamma_n(Q_X))$ is defined as the second term in equation 3.15 and we use the $\gamma_n(Q_X)$ notation to highlight the dependence of γ_n on Q_X. As demonstrated in section 4.3.1, the A-IPCW estimation procedures guarantees that $P_n D^*(Q^*_{X,n}, \Pi_n, \gamma_n) = 0$. If we therefore have that $P_0 R(\Pi_n, \gamma_n)$ converges to 0 if either $\gamma_n(Q_X)$ converges to γ_0 or Π_n is converges to Π_0, then we will have that $P_0 D^F(P^*_{X,n})$ converges to 0.

$$\begin{aligned} P_0 R(\Pi_n, \gamma_n) =& P_0\left(\frac{R - \Pi_n(V)}{\Pi_n(V)}\right)\left(D^F(Q_X)(X) - \gamma_n(Q_X)(V)\right) \\ =& P_0\left(E\left(\left(\frac{R - \Pi_n(V)}{\Pi_n(V)}\right)(D^F(Q_X) - \gamma_n(Q_X)(V))\middle| V\right)\right) \\ =& P_0\left(\left(\frac{E(R|V) - \Pi_n(V)}{\Pi_n(V)}\right)\left(E(D^F(Q_X)|V) - \gamma_n(Q_X)(V)\right)\right) \end{aligned} \tag{3.16}$$

Since $E(R|V) = \Pi_0(V)$ we have that if $\Pi_n \to \Pi_0$ with respect to the $L^2_0(P_0)$ norm, then

$$P_0 R(\Pi_n, \gamma_n(Q_X)) \to 0$$

If instead we have that $\gamma_n(V) \to \gamma_0(V)$, it is clear that

$$\left(E(D^F(Q_X)|V) - \gamma_n(V)\right) = (\gamma_0(V) - \gamma_n(V)) \to 0$$

and the same result will hold.

This result relates to the consistency of the A-IPCW estimator as follows. If the full data gradient itself satisfies an equality

$$P_0 D^F(P^*_{X,n}) = \Psi^F(P_{X,0}) - \Psi^F(P_{X,n}) + R(P^*_{X,n}, P_{X,0})$$

for some term $R(P^*_{X,n}, P_{X,0})$ that converges to 0, than the estimator will be consistent.

For example, the full data gradient for the average treatment effect (equation 3.2) has the property that $R(P^*_{X,n}, P_{X,0})$ converges to 0 if $\bar{Q}_n$ or g_n are consistent for their targets (van der Laan and Rose, 2011). Therefore, an A-IPCW estimator for the average treatment effect will be consistent if either γ_n or Π_n are consistent for their targets *and* either $\bar{Q}_n$ or g_n are are consistent for theirs.

A-IPCW TMLE Implementation

We define and compare three procedures for the generation of A-IPCW TMLE estimators. These correspond with three distinct procedures for the calculation of γ_n, an estimate of $E(D^F(Q_X)|V, R = 1)$. Estimation of this expectation depends on having an estimate of $D^F(Q_X)(X_i)$ for all subjects with $R_i = 1$. This necessitates the estimation of Q_X, components of the full data likelihood P_X such that $D^F(Q_X) = D^F(P_X)$; the specific components are dependent on the form of $D^F(P_X)$. IPCW-TMLE is one approach that can generate these initial estimates, as we describe below.

Initial Estimation of Likelihood Components Generate an initial estimate of Π_n, for example by using main term logistic regression to regress R on V, and denote this estimate $\Pi_{n,1}$. Then perform the IPCW TMLE using the observed data loss function:

$$L(Q_X)(O) = \frac{R}{\Pi_{n,1}(V)} L^F(Q_X)(X)$$

and a submodel of Q_X with parameter ϵ defined such that

$$\frac{\partial}{\partial \epsilon} \log(Q_X(\epsilon)(X))\Big|_{\epsilon=0} = \frac{R}{\Pi(V)} D^F(Q_X)(X)$$

That is, starting with an initial estimator $Q^0_{X,n}$, for $k = 1 \ldots K$, minimize the empirical mean of the loss function of the submodel with respect to ϵ

$$\epsilon^k_n = \underset{\epsilon}{\operatorname{argmin}} P_n L(Q^{k-1}_{X,n}(\epsilon))(X)$$

then set $Q^k_{X,n} = Q^{k-1}_{X,n}(\epsilon^k_n)$. Iterate this procedure until $\epsilon^K_n \approx 0$ and denote the final estimate $Q_{X,n} = Q^{K-1}_{X,n}(\epsilon^K_n)$. For the example parameter of the ATC, this process will result in a set of estimates $Q_{X,n}$ containing g_n, $\bar{Q}_n$ and ψ^F_n. With these components in hand, $D^F(Q_X)(X)$ can be estimated for all subjects with $R_i = 1$.

Augmentation of the Censoring Fit Now calculate $D^F(Q_{X,n})(X)$ for all subjects with $R_i = 1$. Regress $D^F(Q_{X,n})(X)$ on V among these same subjects. This regression will result in a function $\gamma_n(V)$, an estimate of $E(D^F(P_X)|V, R = 1)$, which can be applied

to V for all subjects regardless of their value R. Calculate $\frac{\gamma_n(V)}{\Pi_{n,1}(V)}$ for all subjects. Then perform the updating steps as described in equations 3.11 and 3.12 and iterate until convergence. The final estimate is denoted $\Pi^*_{n,3}$, where the subscript 3 is used to be consistent with the results presented later in the paper.

Implement IPCW-TMLE a Final Time IPCW-TMLE is then performed again, now using a loss function that incorporates the updated missingness fit:

$$L(Q_X)(O) = \frac{R}{\Pi^*_{n,3}(V)} L^F(Q_X)(X). \tag{3.17}$$

This results in $Q^*_{X,n}$, a targeted estimate for $Q_{X,0}$. We also denote this $Q_X(\Pi^*_{n,3})$ to highlight its dependence on the specific missingness fit.

Final Estimator $Q^*_{X,n}$, is used with the full data mapping to generate the final estimator $\psi_{n,3} = \Psi^F(Q^*_{X,n})$.

We also consider an extension to this procedure, in which the missingness fit and the corresponding likelihood components, $Q_X(\Pi_n)$, are both iteratively updated. That is, this process uses each iteration of Π^k_n in the IPCW TMLE procedure, to estimate $Q_X(\Pi^k_n)$. Given $Q_X(\Pi^k_n)$ $D^F(Q_X(\Pi^k_n))(X)$ is calculated among subjects with $R_i = 1$ and regressed on V_i to generate $\gamma^k_n(V)$. A new parametric submodel is defined, which incorporates this new estimate of γ_0:

$$\Pi^k_n(\beta)(V) = \text{expit}\left(\text{logit}(\Pi^k_n(V)) + \beta\left(\frac{\gamma^k_n(V)}{\Pi^k_n(V)}\right)\right).$$

The empirical mean of the logistic loss function is minimized with respect to β over the observed data

$$\beta^k_n = \underset{\beta}{\text{argmin}}\, P_n L(\Pi^{k-1}_n(\beta))(O)$$

for $k = 1, \ldots, K$. This is iterated until $\beta^k_n \approx 0$ and $\Pi^{k-1}_n(\beta^k_n) \equiv \Pi^*_{n,4}$. IPCW-TMLE is then implemented a final time, using $\Pi^*_{n,4}$ and Q^*_X, the resulting targeted estimate of $Q_{X,0}$, generates $\psi_{n,4} = \Psi^F(P^*_X)$.

A third approach that we consider is the use of multiple imputation to estimate the parameters of the regression, γ_n. That is, values $L^{(m)} = (L^{(m)}_1, \ldots, L^{(m)}_j)$ for $m = 1 \ldots M$ are imputed for the j subjects where $R_i = 1$. $L^{(m)}_i$ are combined with $\{V_i : R_i = 0\}$ and $\{(V_i, L_i) : R_i = 1\}$ to create M $\bar{X}^{(m)}$s, imputed full data sets. For each $\bar{X}^{(m)}$, the full data mapping $\hat{\Psi}^F(\bar{X}^{(m)})$ can be used to estimate $Q^{(m)}_{X,n}$ and therefore calculate $D^F(Q^{(m)}_{X,n})(X_i)$ for all subjects. A parameterization of the regression of $D^F(Q)(X)$ on V, γ_n, indexed by χ, is assumed, and $\chi^{(m)}_n$ is estimated using each of the data sets. Let $\chi_n = \frac{1}{M}\sum_m \chi^{(m)}_n$ and $\gamma_{n,5} = \gamma_n(\chi_n)$ be the estimate of $E(D^F(P_X)|V, R = 1)$. A new parametric submodel

$\{\Pi_n^k(\beta) : \beta\}$ is defined such that:

$$\Pi_n^k(\beta)(V) = \text{expit}\left(\text{logit}(\Pi_n^k(V)) + \beta\left(\frac{\gamma_{n,5}(V)}{\Pi_n^k(V)}\right)\right).$$

and iteratively updated as described in equation 3.12 until $\beta_n^k \approx 0$. $\Pi_{n,5}^* = \Pi_{n,1}^{K-1}(\beta_n^K)$ is the final estimator for the missingness mechanism and a corresponding loss function, as in equation 3.17, is defined. Q_X^* is the resulting targeted estimate for $Q_{X,0}$ and $\psi_{n,5} = \Psi^F(Q_X^*)$ is then the A-IPCW TMLE for ψ_0.

Variance Estimation

For all the procedures described, the asymptotic variance of $\sqrt{n}(\psi_n - \psi_0^F)$ can be estimated by calculating the empirical variance of $D^*(Q_X^*, \Pi_n^*, \gamma_n)$. Since $P_n D^* = 0$, this is equivalent to the mean square of the empirical influence curve and

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} D^*(Q_X^*, \Pi_n^*, \gamma_n)(O_i)^2.$$

95% confidence intervals can then be calculated as $\psi_n \pm 1.96\frac{\sigma_n}{\sqrt{n}}$. For the ATE estimatand, this variance estimator is conservative if Π_n and g_n are consistent for their targets, and correct if γ_n and Q_n are also consistent.

3.5 Simulation

We examined the performance of the proposed estimators in simulation. Two simulation studies were run. The first explored the performance of the various estimators as the proportion of missingness changed while the second explored performance under misspecification of Π and γ. The data structure contained the following variables whose variable definitions and relationships mimic the motivating data set presented in the applied example in section 5. A directed acyclic graph representing the relationships between the variables is contained in figure 3.1.

W (Age, Sex, Race) where Age $\sim N(46, \sqrt(103))$, Sex $\sim$ Bern(0.7), Race $\sim$ Bern(0.7) are independent from each other.

L (Smoking, BMI) where Smoking $\sim$ Bern(logit($\beta_0^l + \beta_1^l W$)) with $\beta_0^l = -1.8$ and $\beta_0^l = (0.4, .2, .18)$. BMI $\sim N(\mu_l, \sqrt{28})$ where $\mu_l = \beta_2^L + \beta_3^L W$ and $\beta_2^l = 28, \beta_3^L = (.03, .95, .67)$

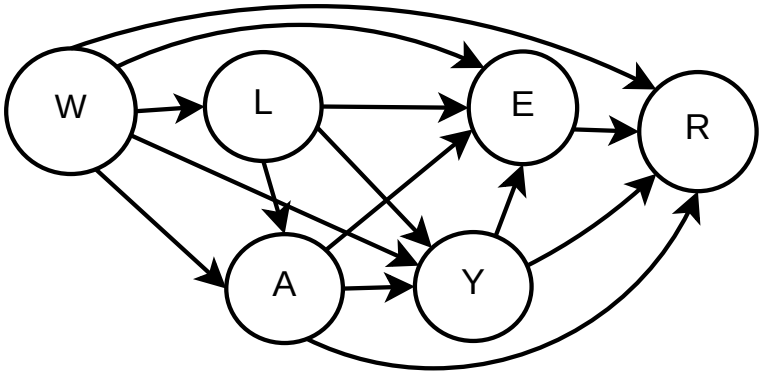

Figure 3.1: Directed Acyclic Graph (DAG) of variables used in simulation study

A (Exposure) where $A \sim \text{Bern}(\text{logit}(\beta_0^A + \beta_1^A W + \beta_2^A L))$ and $\beta_0^A = 2.2, \beta_1^A = (-.02, -.9, .37)$, $\beta_2^A = (-.04, -.11)$.

Y (Outcome) where $Y \sim \text{Bern}(\text{logit}(\beta_0^Y + \beta_1^Y W + \beta_2^Y L + \beta_3^Y A))$, and $\beta_0^Y = -8.7, \beta_1^Y = (.12, .21, -.19), \beta_2^Y = (.05, .49), \beta_3^Y = .2$.

E (Leaving Work) where $E \sim \text{Bern}(\text{logit}(\beta_0^E + \beta_1^E W + \beta_2^E L + \beta_3^E A + \beta_4^E Y)$, and $\beta_0^E = -3.0, \beta_1^E = (.04, -.06, -.03), \beta_2^E = (.006, .44), \beta_3^E = .2, \beta_4^E = -.5$.

R (Measurement of L) where $R \sim \text{Bern}(\text{logit}(\beta_0^R + \beta_1^R W + \beta_3^R A + \beta_4^R Y + \beta_5^R E))$, and $\beta_0^R = -2, \beta_1^R = (.04, 0, 0), \beta_3^R = 1, \beta_4^R = -.5, \beta_5^R = -.6$.

The parameter of interest used for the study was the same as our running example in the paper, the ATE of the treatment A: $\psi^F = \Psi^F(P_X) = E_{W,L}(\bar{Q}(Y|A = 1, L, W) - \bar{Q}(Y|A = 0, L, W))$. This parameter has the following efficient influence curve, where we define $g(a|W, L) = P(A = a|W, L)$:

$$\begin{aligned} D^F(P_X)(X) = \quad & \left(\tfrac{I(A=1)}{g(1|W,L)} - \tfrac{I(A=0)}{g(0|W,L)}\right)(Y - \bar{Q}(W, L, A)) \\ & + \bar{Q}(W, L|A = 1) - \bar{Q}(W, L|A = 0) - \psi^F. \end{aligned}$$

The components of P_X that need to be estimated in order to calculate $D^F(P_X)$ are $Q_X = (\bar{Q}, g, P_{W,L}, \psi^F)$ and $D^F(P_X) = D^F(Q_X)$. The full data mapping procedure we used was simple TMLE for the average treatment effect as fully defined in,van der Laan and Rubin (2006a) or Moore and van der Laan (2007).

Five estimators $(\psi_{n,1}, \psi_{n,2}, \psi_{n,3}, \psi_{n,4}, \psi_{n,5})$ were calculated in each of the simulated data sets. $\psi_{n,1}$ was traditional IPCW-TMLE, $\psi_{n,3}$ was the basic augmented IPCW-TMLE, $\psi_{n,4}$ was the augmented TMLE with an iterative update of $Q_{X,n}$, and $\psi_{n,5}$ was an augmented IPCW-TMLE where γ_n was estimated from multiply imputed data sets. We also included a multiple imputation type estimator, $\psi_{n,2}$. For this, the full data estimation procedure

was performed in each of M = 5 multiply imputed data sets and the results were combined together using Rubin's rules. All analysis was performed in R 3.0 (R Core Team, 2013a) and the package `mi` (S Su et al., 2011) was used to perform the multiple imputation.

The first simulation was run with n = 15,000 and compared the performances of the various estimators when each of the estimator components were correctly specified, but the proportion of missingness ($E_{P_X}(R)$) was changed. The intercept of the regression predicting R, β_0^R, was set to 9 different values in order to vary this proportion. For each intercept value, 100 data sets were created and the average bias, mean squared error (MSE) and coverage probabilities of the estimators were compared.

Table 3.5 contains the results from this simulation which demonstrate that under correct model specification, all five estimators are approximately unbiased. The multiple imputation based estimator $\psi_{n,2}$ delivered the lowest mean squared error of the five estimators. The performance of the augmented-IPCW TMLE estimators ($\psi_{n,3}$, $\psi_{n,4}$, and $\psi_{n,5}$) demonstrated slight reductions in bias and MSE as well as increased coverage probabilities, as compared to IPCW TMLE ($\psi_{n,1}$). The coverage probabilities for the augmented estimators were consistently higher than 95% and increased as the proportion of missing data increased. This can be attributed to the fact that the true efficient influence curve for the augmented estimator is D^* minus its projection onto the space of scores for the missingness mechanism (van der Laan and Robins, 2003, Tsiatis, 2006). Therefore, the variance estimates we use here, which are based on D^* only, are guaranteed to be conservative and the coverage probability is greater than 95%.

Figure 3.2 demonstrates that as the proportion of missingness within the population goes up, the gains due to the augmentation procedure become larger. It plots the ratio of the relative mean squared error of the simple augmented IPCW-TMLE($\psi_{n,3}$) to that of the IPCW-TMLE ($\psi_{n,1}$) as a function of $E_{P_X}(R)$. The augmentation procedure results in increased efficiency as the proportion of missingness increases. The efficiency gains are modest at relatively low missingness proportions, but become substantial as this proportion of missingness moves above 40%.

The second simulation explored the performance of the augmented estimators under misspecification of the Π and γ estimator components. The misspecification of Π was performed by generating models, Π_n that did not include the members of W `Age` or `Sex`. γ was misspecified by generating models for $E(D^F(P_X)|V)$ that also did not include these predictors. `Age` or `Sex` were also not passed to the multiple imputation procedure when calculating $\psi_{n,2}$ and $\psi_{n,5}$ under misspecification of γ. Table 3.5 contains the MSEs and biases for the five estimators calculated under these conditions on 100 data sets with $n = 5000$.

The A-IPCW estimators $\psi_{n,3}$ and $\psi_{n,4}$ outperform the IPCW estimator $\psi_{n,1}$ under each of the study conditions. $\psi_{n,5}$ had lower absolute bias under each condition, but had a higher MSE under misspecification of Π. When γ is correctly specified, the multiple imputation

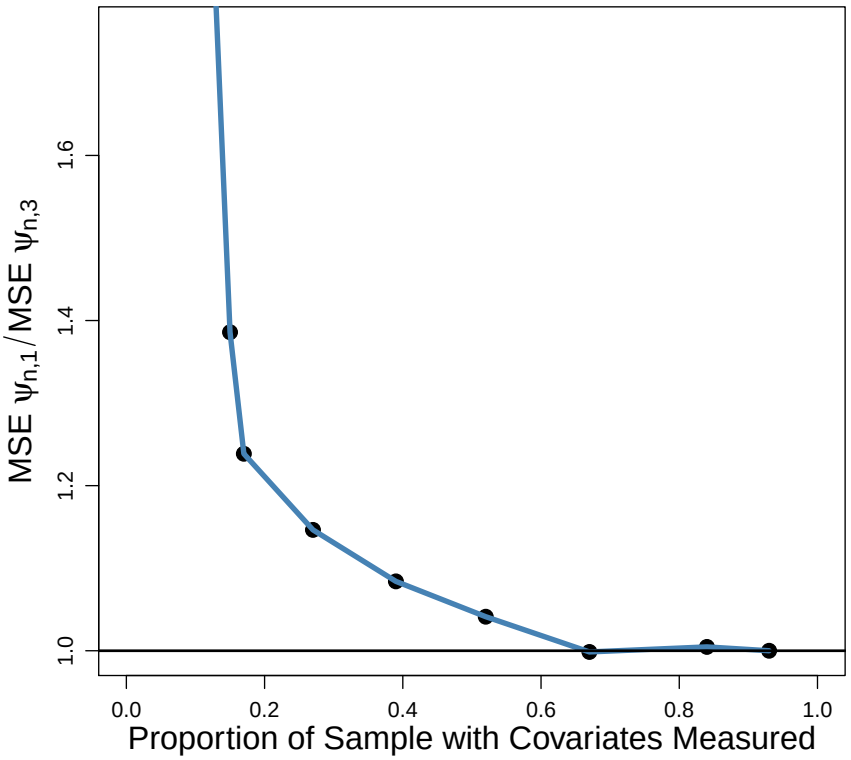

Figure 3.2: Relative efficiency of augmented IPCW-TMLE ($\psi_{n,3}$) compared to simple IPCW-TMLE ($\psi_{n,1}$)

estimator $\psi_{n,2}$ exhibited lower absolute bias and MSE than any of the inverse weighted estimators. When γ is misspecified, however, $\psi_{n,2}$ demonstrates bias which is not present in the IPCW estimators.

Figure 3.3 shows the distribution of the bias for the five estimators among the 100 data sets under each simulation condition. Under correct specification of γ and Π, each of estimators are unbiased. When Π is misspecified, however, the IPCW estimator $\psi_{n,1}$ is biased, while the MI estimator $\psi_{n,2}$ and the A-IPCW estimators $\psi_{n,3}$ and $\psi_{n,4}$ are unbiased. This occurs despite the fact that the estimation of $D^F(Q_{X,n}(\Pi_n))$ is dependent on a misspecified Π_n for the A-IPCW estimators. We also note that both the bias and the variance for $\psi_{n,5}$ under misspecification of Π_n is higher than expected. This may occur because we are not assured of fully correct specification of γ_n, even when all relevant variables are included as predictors, because the true form of the conditional regression is unknown. It is also possible that under the simulation conditions, the second order bias becomes large (Kang and Schafer, 2007). Future work with these estimators will explore the application of flexible modeling approaches for γ which may improve the performance of A-IPCW TMLE estimators of the sort described here.

3.6 The Aluminum Worker Cohort

The present course of work was motivated by research on an occupational cohort of aluminum smelter workers, in which we were interested in studying the effects of exposure to airborne particulate matter with an aerodynamic diameter of less than 2.5 μm ($PM_{2.5}$) on heart disease incidence. The target paramter was the contrast between incidence of ischemic heart disease (IHD) among workers constantly exposed above or below a cut-off for 15 years of work. The cohort contained 5,426 workers who were followed for a maximum of 15 years between 1996 and 2013, or until they left work. Worker's occupational exposure to $PM_{2.5}$ and experience of ischemic heart disease (IHD) were measured annually during follow-up, following a two year washout period designed to remove prevalent cases from the cohort. Two important potential confounders were body mass index (bmi) and smoking status, which were incompletely measured in the population, with 3,914 (62%) workers having this information recorded. This cohort and the application of longitudinal TMLE of a mean outcome to it have been fully characterized by Costello et al. (2014) and in chapter 4.1 of this dissertation. This applied example represents one of the several target parameters we consider therein. We describe the application of the augmented IPCW TMLE to this analysis in the section that follows.

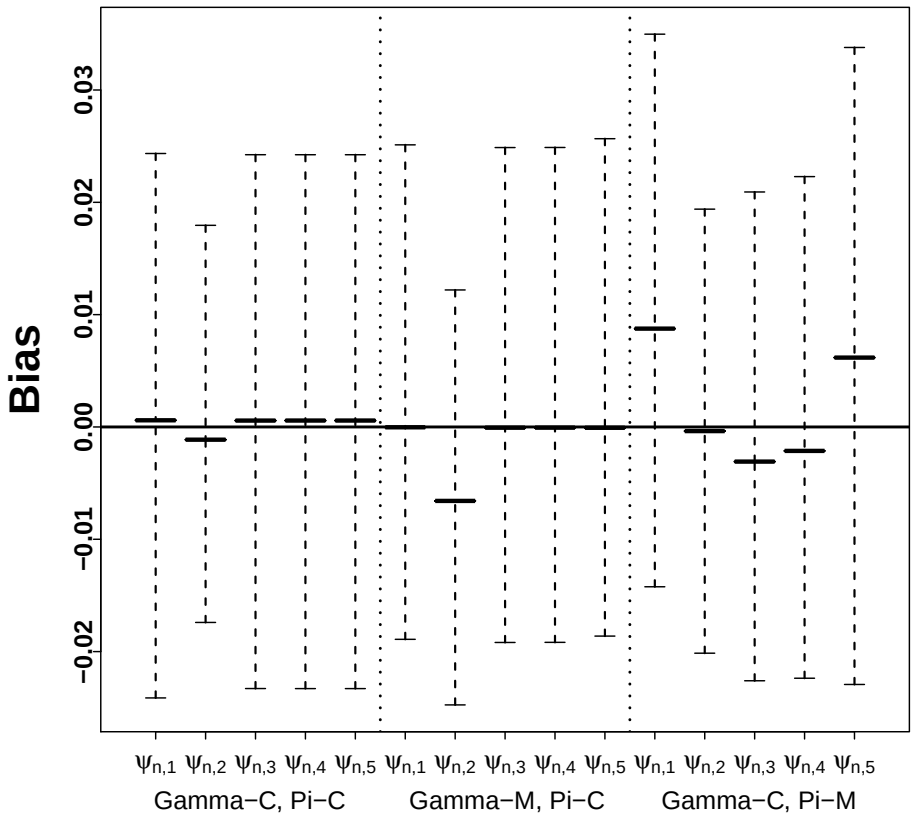

Figure 3.3: Distribution of bias for the 5 estimators $\psi_{n,1}, \psi_{n,2}, \psi_{n,3}, \psi_{n,4}$ and $\psi_{n,5}$ implemented using correctly and incorrectly specified components Π and γ. $\psi_{n,1}$ is the IPCW TMLE, $\psi_{n,2}$ is the full data TMLE averaged over 5 multiply imputed data sets, and $\psi_{n,3}$, $\psi_{n,4}$ and $\psi_{n,5}$ are alternative approaches to implementing A-IPCW TMLE. Simulation condition Π-M γ-C indicates that the probability of missingness $\Pi = P(R|V)$ was misspecified while the conditional regression of the full data influence curve $\gamma = E(D^F|V)$ was correctly specified (in that it contained all relevant members of V)

3.6.1 Data Structure and Target Parameter

The full data structure for the aluminum smelter worker cohort was:

$$X = (W(0), L, A(1), W(1), \ldots, A(15), Y \in W(15), R)$$

with variables defined as follows:

W(t) Completely recorded variables, both time variant and invariant, as measured at end of time point t. These include age, race, facility location, marital status, job grade, calendar year, prior diagnosis of ischemic heart disease, diabetes, hypertension, dyslipidemia, or clinical obesity, risk score (an insurance-based health index), time off work, time since hire, and an indicator of leaving work prior to 2009.

L Incompletely recorded time-invariant variables (bmi and smoking status).

A(t) A vector of intervention variables containing: $E(t)$ an indicator of employment during time t at a job where exposure to $PM_{2.5}$ has been estimated to be above $1.77\frac{mg}{m^3}$ and $C(t)$ an indicator of either active work status while younger than 55,or being older than 55 at the end of time point t.

Y(t) An indicator of diagnosis with incident ischemic heart disease by the end of time point t. $Y \equiv Y(15)$ is a measure of cumulative incidence by the end of follow-up and is the variable about which we wish to make inference.

R An indicator of recording of L.

Measurement of BMI and smoking status were acquired from two sources: 1) Review of on-site medical records performed by study personnel in 2009 and 2) A worker health database maintained by the employer. A primary determinant of whether the medical records were available for review was whether a worker was actively employed at the time. The two sources were thus necessary to believe that the positivity assumption could hold for workers who were not employed in 2009. Information on work termination was consequently included in $W \subset V$ to ensure that the MAR assumption was met.

The observed data structure was:

$$O = (W(0), LR, A(1), W(1) \ldots A(15), Y \in W(15), R)$$

where we set all variables at time points after a worker leaves work to their last recorded values. That is, if a worker leaves at time point k, then $W(t) = W(k)$ and $A(t) = A(k)$ for

$t = k, \ldots, K$. We can express the likelihood as

$$\begin{aligned} P_0(O) &= \Big\{ P_0(L|Pa(L)) \prod_{k=0}^{K} P_0(W(k)|Pa(W(k))) \prod_{k=1}^{K} P_0(A(k)|Pa(A(k))) \Big\}^R \\ &\quad \Big\{ \int_l \prod_{k=0}^{K} P_0(W(k)|Pa(W(k))) \prod_{k=1}^{K} P_0(A(k)|Pa(A(k))) \partial \nu_L \Big\}^{1-R} \\ &= \Big\{ P_0(L|Pa(L)) \prod_{k=0}^{K} Q_{0,W(k)}(O) \prod_{k=1}^{K} g_{0;A(k)}(O) \Big\}^R \\ &\quad \Big\{ \int_l \prod_{k=0}^{K} Q_{0,W(k)}(O,l) \prod_{k=1}^{K} g_{0;A(k)}(O,l) \partial \nu_L \Big\}^{1-R} \end{aligned}$$

where $Pa(W(k))$, $Pa(A(k))$ and $Pa(L)$ denote the parents (all variables that directly affect the values) of L of $W(k)$, $A(k)$, and L, respectively. $Q_{0,W(k)}$ is the true conditional distribution of $W(k)$ given its parents and $g_{0,A(k)}$ is the true conditional distribution of $A(k)$ given its parents. We also use the notation $g_{0:k} \equiv \prod_{j=0}^{k} g_{A(j)}$. We define a statistical model $\mathcal{M}$ for P_0 as $\mathcal{M} = \{P = P_L Q g, Q \in \mathcal{Q}, g \in \mathcal{G}\}$ where $\mathcal{Q}$ contains all possible values for $Q_{0,W(k)} : k = 0 \ldots K$ and $\mathcal{G}$ contains all possible values for $g_{0,A(k)} : k = 1 \ldots K$.

Our target parameter of interest is the cumulative incidence from ischemic heart disease among workers following specified intervention regimes, $\bar{d}$ that specify the values of the intervention nodes A at each time point. The two regimens of interest correspond to workers always being exposed either above or below the exposure cut-off while remaining at work until retirement age, after which they are free to leave as they wish. That is, we define two treatment regimens corresponding to two levels of the binary exposure, $d_1(\bar{W})$ and $d_0(\bar{W})$ such that

$$\begin{aligned} d_e(\bar{W}(t))(t+1) &= (e,1) \text{ if age}(t) < 55 \\ d_e(\bar{W}(t))(t+1) &= (e,0) \text{ if age}(t) > 55 \end{aligned}$$

Let $P^d(w) = \prod_{k=0}^{K} Q^d_{W(k)}(\bar{w}(k))$ be the conditional distribution of w under the intervention $A(t) = d_e(\bar{W}(t))$ for $t = 1, \ldots K$, which we denote with $\bar{A}(K) = \bar{d}(K)$ for notational clarity. Here $Q^d_{W(k)}(w(k)) = Q_{W(k)}(w(k)|\bar{W}(k-1), \bar{A}(k) = \bar{d}(k))$. Then let $W^d \sim P^d$ be the distribution of the covariates under intervention regimen $\bar{d}$ and let Y^d be its final element. We are interested in making inferences about the parameters of the distribution of this variable under different interventions. Our target parameter is the mean of Y^d, $\psi_0 = E_{P^d} Y^d$. Our full data procedure (what we would have applied to the full data structure if L was measured) was longitudinal TMLE of a mean outcome procedure, which is a mapping from the model to the reals, $\Psi : \mathcal{M} \to \mathbb{R}$ such that $\Psi(P) = E_{P^d} Y^d$.

3.6.2 Longitudinal TMLE of a mean outcome

The full data procedure we used to estimate $E_{P^d}Y^d$ was longitudinal TMLE of a mean outcome. This method involves representing the target parameter as an iterative conditional expectation, as first demonstrated by Robins (2000a) and further explicated by Bang and Robins (2005). By the tower rule of conditional expectations:

$$EY^d(K+1) = E(\ldots E(E(Y|\bar{W}(K), \bar{A}(K) = \bar{d}(K))|\bar{W}(K-1), \bar{A}(K-1) = \bar{d}(K-1))\ldots|W(0)).$$

The mean of the $Y^d(K+1)$ can therefore be estimated by conditioning first on $\bar{W}(K), \bar{A}(K) = \bar{d}(K)$, then on $\bar{W}(K-1), \bar{A}(K-1) = \bar{d}(K-1)$ and so on until $W(0)$. Longitudinal TMLE of a mean outcome was first described by Van der Laan and Gruber in 2012 (van der Laan and Gruber, 2012), and is a method for estimating estimands like $E_{P^d}Y^d$ using targeted fits of each of these conditional expectations. Each fit is targeted to ensure that the final estimator solves the efficient influence curve equation. The efficient influence curve for this estimator in $\mathcal{M}$ can be written as $D^F = \sum_{k=1}^{K+1} D_k^F$ where:

$$\begin{aligned} D_{K+1}^F &= \frac{I(\bar{A}(K) = \bar{d}(K))}{g_{0:k}}(Y - \bar{Q}^d_{W(K+1)}) \\ D_k^F &= \frac{I(\bar{A}(k-1) = \bar{d}(k-1))}{g_{0:k-1}}(\bar{Q}^d_{W(k+1)} - \bar{Q}^d_{W(k)}) \\ D_0^* &= \bar{Q}^d_{W(1)} - \psi_0 \end{aligned}$$

$\bar{Q}^d_{W(K+1)} = E(Y|\bar{W}(K), \bar{A}(K) = \bar{d}(K))$ and $\bar{Q}^d_{W(k)} = E(\bar{Q}^d_{W(k+1)}|\bar{W}(k-1), \bar{A}(k-1) = \bar{d}(k-1))$. Note that $D^F(P_X)$ is a function of P_X through the likelihood components $Q_X = (\bar{Q}^d_{W(1)}, \ldots \bar{Q}^d_{W(15)}, g_{0:15})$. The longitudinal TMLE of a mean outcome procedure generates targeted fits, $\bar{Q}^*_{X,n}$, of Q_X that ensure that the empirical sum of the efficient influence curve will be 0:

$$\frac{1}{n}\sum_{i=1}^{n} D^F(\bar{Q}^*_{X,n}, g_{n,0:15})(O_i) = 0$$

This is a necessary condition of the full data estimation procedure in order to be able to apply AIPCW-TMLE. We used a modified version of the `ltmle` package (Schwab et al., 2013, Lendle et al., 2014) and R 3.0.3 to perform our analysis.

3.7 Implementation of A-IPCW TMLE

The first step of AIPCW-TMLE implementation is to generate estimates of the necessary likelihood components, Q_X, using IPCW-TMLE. We used logistic regression to create an initial estimate of the missingness mechanism, $\Pi_{n,1}(V)$, by regressing R on $V = (W(15), A(15))$.

Weights $p_i = (\Pi_{n,1}(V))^{-1}$ were then calculated for each worker. These were used in the next step to estimate $g_{n,0:K}$ with weighted logistic regression models. The A vector has two components, one for the exposure E and the other for the censorship state, C. We could therefore partition g_0 into two corresponding distributions

$$\begin{aligned} g_{0,E(t)} &= P_0(E(t)|\bar{W}(t), L, C(t-1)=1, Y(t-1)=0) \\ g_{0,C(t)} &= I(\text{Age(t) } >55) + P_0(C(t)|\bar{W}(t), L, E(t), C(t-1)=1, Y(t-1)=0, \text{Age(t) } <55) \\ g_{0,A(t)} &= g_{0,E(t)} g_{0,C(t)} \end{aligned}$$

Estimates of $g_{0,A(t)}$ were made by fitting logistic regressions weighted by p_i among workers with $R_i = 1$. We made a Markov assumption that the intervention nodes at time t were only a function of the parent nodes as measured at time $t-1$. This allowed us to pool observations over time and generate a single fit for each distribution, $g_{n,E}(W(t-1), L, E(t-1), t, C(t-1) = 1, Y(t-1) = 0)$ and $g_{n,C}(W(t-1), L, E(t), t, C(t-1) = 1, Y(t-1) = 0,$ Age(t) $<55)$. We then defined:

$$\begin{aligned} g_{n,0:K} \equiv \prod_{t=0}^{K} & \quad g_{n,E}(E(t)|W(t-1), L, t, E(t-1), C(t-1)=1, Y(t-1)=0) \\ & g_{n,C}(C(t)|W(t-1), L, t, E(t), C(t-1)=1, Y(t-1)=0, \text{Age } <55) \end{aligned}$$

which we used in the targeting of the sequential regressions as described below.

We generated an estimate for the first conditional expectation, $\bar{Q}^d_{Y,n}$ of $\bar{Q}^d_{Y,0} = E_0(Y^d|\bar{W}^d(15), L, \bar{A}(15) = \bar{d}(15))$ by fitting a weighted logistic regression of Y on $(W(15), L, \bar{A}(15) = \bar{d}(15), Y(14) = 0, R = 1)$, with weights p. The Markov assumption that only the members of W as measured at time point $t-1$ affected the distribution of Y at time t was made for this and all subsequent conditional regressions. We then updated this fit by regressing Y on $(I(\bar{A}(15) = \bar{d}(15)))(g_{n,0:15}(\bar{W}(15), L))^{-1}$, using $\bar{Q}^d_{Y,n}$ as an offset and weights p. The result was the final targeted fit of the conditional regression, $\bar{Q}^{d,*}_{Y,n}$.

We then repeated a similar process for time point $t = 14$. We first generated $\bar{Q}^d_{W(14),n}$ by regressing $\bar{Q}^{d,*}_{Y,n}(W(15), L)$ on $(W(14), L, \bar{A}(14) = \bar{d}(14), Y(13) = 0, R = 1)$. We then updated this fit, generating $\bar{Q}^{d,*}_{W(14),n}$, by regressing $\bar{Q}^{d,*}_{Y,n}$ on $(I(\bar{A}(14) = \bar{d}(14)))(g_{n,0:14}(\bar{W}(14), L))^{-1}$ on with offset $\bar{Q}^d_{W(14),n}$. This process was repeated 13 times until arriving at a final regression $\bar{Q}^{d,*}_{W(1),n}$. The final estimator was this function applied to the weighted empirical distribution of fully observed baseline covariates.

$$\psi_n = \frac{1}{\sum_{i=1}^n \frac{R_i}{p_i}} \sum_{i=1}^n \frac{R_i}{p_i} \bar{Q}^{d,*}_{W(1),n}(W_i(0), L_i)$$

This process resulted in an estimate $Q_{X,n} = (\bar{Q}^{d,*}_{Y,n}, \ldots, \bar{Q}^{d,*}_{W(1),n}, g_{n,0:15})$, which could be

plugged into the formula for the full data efficient influence curve $D^F(Q_{X,n})(X)$.

The application of augmented IPCW TMLE procedure was straightforward at this point. $D^F(Q_{X,n})(X)$ was regressed on $V = (A(0), A(15), W(0), W(15))$ among subjects with $R_i = 1$. This resulted in $\gamma_n(V)$, an estimate of the regression of the full data efficient influence curve for the longitudinal data structure on the always observed variables. $\gamma_n(V)$ was then used to update the fit of the missingness mechanism by regression of $(\gamma_n(V_i))(\Pi_n(V_i))^{-1}$ on R_i, with offset $\Pi_n^{(0)}(V_i)$. After iteration, we arrived at a final fit $\Pi_n^*(V)$ and new weights $p^* = (\Pi_n^*(V))^{-1}$. The new weights were used to repeat the process described above and the final estimator was $\psi_{n,3}$. The other augmented IPCW TMLE estimators $\psi_{n,4}$ and $\psi_{n,5}$ were implemented in a similar manner, excepting for the previously described differences in the estimation of γ.

3.7.1 Results

We applied the five estimation procedures to the aluminum smelter cohort and estimated the counterfactual cumulative incidence of IHD under two different intervention regimens. These regimens involve assignment of an exposure level as well as the prevention of censoring due to leaving work prior to retirement age of 55. That is, censoring is defined within this data structure leaving work when younger than 55 and the regimens of interest included an intervention to prevent censoring. Workers following the first regimen ($\bar{e} = \bar{1}$) work in jobs where they are exposed to greater than 1.77 $\frac{mg}{m^3}$ $PM_{2.5}$ for the duration of their work experience and do not leave work when younger than 55. The experience of workers following the first regimen is compared to that of workers following a second exposure regimen ($\bar{e} = \bar{0}$) in which the intervention on censoring is maintained, but exposures are all at levels lower than the cut-off of 1.77 $\frac{mg}{m^3}$ $PM_{2.5}$. Workers older than 55 were allowed to leave work at the same time as they did under their observed exposure history, under the assumption that retirement choices in this population are less likely to be affected by exposure and health history. The definition of intervened-on censoring was made to prevent potential selection bias due to unhealthy workers leaving work while maintaining a realistic treatment regimen to ensure identifiability of the target estimator. (van der Laan and Petersen, 2007, Bembom and van der Laan, 2007)

Table 3.7.1 contains the results of this estimation. Each method returned similar estimates of approximately .076 for the cumulative incidence of IHD at year 15 under the exposed intervention regimen ($Y^{\bar{1}}(15)$). The inverse weighted estimators ($\psi_{n,1}, \psi_{n,3}, \psi_{n,4}, \psi_{n,5}$) all estimated similar incidences of 0.044 and 0.043 for the unexposed intervention regimen ($Y^{\bar{0}}(15)$). The multiple imputation estimator ($\psi_{n,2}$), however, returned a higher point estimate of 0.054. The resulting rate ratios reflect these differences in point estimates. The augmented-IPCW estimators have slightly narrower confidence bands around them compared to the IPCW-TMLE estimator, and slightly wider intervals than the MI estimator.

The narrower confidence bands generated by the augmented estimators result in the statistical significance of their rate ratios, with p-values of 0.02, while the traditional IPCW estimator returned a p-value of 0.08. There was virtually no difference in performance between the three approaches to implementing the augmented IPCW estimator.

	$\psi_{n,1}$			$\psi_{n,2}$			$\psi_{n,3}$			$\psi_{n,4}$			$\psi_{n,5}$		
$P(\Pi = 1)$	**Bias**	**MSE**	**CProb**	**Bias**	**MSE**	**CProb**	**Bias**	**MSE**	**CProb**	**Bias**	**MSE**	**CProb**	**Bias**	**MSE**	**CProb**
0.93	-8.57e-04	3.45e-05	0.95	-7.44e-04	3.31e-05	0.97	-8.57e-04	3.45e-05	0.95	-8.57e-04	3.45e-05	0.95	-8.58e-04	3.45e-05	0.95
0.84	-2.21e-04	4.68e-05	0.94	-2.97e-04	4.41e-05	0.92	-2.24e-04	4.66e-05	0.94	-2.24e-04	4.66e-05	0.94	-2.13e-04	4.69e-05	0.94
0.67	-6.18e-04	4.93e-05	0.97	-5.17e-04	4.58e-05	0.94	-6.01e-04	4.93e-05	0.97	-6.01e-04	4.93e-05	0.97	-6.13e-04	4.92e-05	0.97
0.52	3.42e-04	5.52e-05	0.97	7.62e-05	3.53e-05	0.98	3.23e-04	5.30e-05	0.97	3.24e-04	5.30e-05	0.97	4.01e-04	5.44e-05	0.98
0.39	5.38e-04	5.84e-05	0.97	1.26e-04	3.73e-05	0.94	5.41e-04	5.38e-05	0.99	5.44e-04	5.38e-05	0.99	5.74e-04	5.71e-05	0.98
0.27	-6.03e-04	1.24e-04	0.95	-4.80e-04	4.23e-05	0.93	-5.58e-04	1.08e-04	0.99	-5.56e-04	1.08e-04	0.99	-5.92e-04	1.16e-04	0.96
0.17	1.19e-04	1.98e-04	0.91	-5.61e-05	5.55e-05	0.88	9.76e-04	1.60e-04	1.00	9.95e-04	1.61e-04	1.00	1.07e-03	1.76e-04	0.92
0.15	-3.12e-03	1.88e-04	0.95	2.00e-04	4.55e-05	0.97	-1.01e-03	1.35e-04	1.00	-9.94e-04	1.36e-04	1.00	-9.99e-04	1.53e-04	0.98
0.12	-9.25e-03	3.68e-04	0.90	-2.80e-04	3.66e-05	0.96	6.49e-04	1.86e-04	1.00	7.44e-04	1.88e-04	1.00	8.49e-04	2.17e-04	0.98
0.09	-2.14e-02	8.66e-04	0.79	-2.86e-04	5.16e-05	0.94	2.35e-03	2.35e-04	1.00	2.50e-03	2.37e-04	1.00	3.26e-03	2.97e-04	0.99

Table 3.1: Bias, mean squared error (MSE) and coverage probability (CProb) for the five estimators in simulation in simulation as the percentage of missing data is varied. $\psi_{n,1}$ is traditional IPCW-TMLE; $\psi_{n,2}$ is a multiple imputation type estimator; $\psi_{n,3}$ is a basic augmented IPCW-TMLE; $\psi_{n,4}$ is an augmented IPCW-TMLE with an iterative update of $Q_{X,n}$; $\psi_{n,5}$ is an augmented IPCW-TMLE where γ_n was estimated from multiply imputed data sets.

Simulation	$\psi_{n,1}$		$\psi_{n,2}$		$\psi_{n,3}$		$\psi_{n,4}$		$\psi_{n,5}$	
Condition	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**
Π-C γ-C	5.93e-04	2.27e-04	-1.13e-03	1.12e-04	5.68e-04	2.15e-04	5.72e-04	2.15e-04	5.76e-04	2.24e-04
Π-M γ-C	-3.03e-05	1.82e-04	-6.59e-03	1.75e-04	-6.56e-05	1.77e-04	-6.32e-05	1.77e-04	-7.33e-05	1.82e-04
Π-C γ-M	8.75e-03	3.06e-04	-3.69e-04	1.38e-04	-3.08e-03	1.95e-04	-2.13e-03	1.92e-04	6.18e-03	3.62e-04

Table 3.2: Bias and mean squared error (MSE) for the five estimators under correct specification and misspecification of the estimator components. Simulation condition Π-M γ-C indicates that the probability of missingness $\Pi = P(R|V)$ was misspecified while the conditional regression of the full data influence curve $\gamma = E(D^F|V)$ was correctly specified (in that it contained all relevant members of V). $\psi_{n,1}$ is traditional IPCW-TMLE; $\psi_{n,2}$ is a multiple imputation type estimator; $\psi_{n,3}$ is a basic augmented IPCW-TMLE; $\psi_{n,4}$ is an augmented IPCW-TMLE with an iterative update of $Q_{X,n}$; $\psi_{n,5}$ is an augmented IPCW-TMLE where γ_n was estimated from multiply imputed data sets.

	$Y_{\bar{1}}(15)$		$Y_{\bar{0}}(15)$		Rate Ratio		
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	p-value
$\psi_{n,1}$	0.077	(.041, .112)	0.044	(.025, .063)	1.75	(0.93, 3.31)	0.08
$\psi_{n,2}$	0.075	(.048, .101)	0.054	(.032, .076)	1.38	(0.80, 2.37)	0.25
$\psi_{n,3}$	0.076	(.049, .103)	0.043	(.028, .058)	1.76	(1.10, 2.84)	0.02
$\psi_{n,4}$	0.076	(.049, .103)	0.043	(.028, .058)	1.76	(1.10, 2.84)	0.02
$\psi_{n,5}$	0.076	(.049, .103)	0.043	(.028, .058)	1.76	(1.10, 2.84)	0.02

Table 3.3: The five estimators as applied to the aluminum smelter worker cohort, comparing the cumulative incidence of ischemic heart disease at 15 years among workers exposed to two different $PM_{2.5}$ exposure and censoring regimens. $\bar{a} = \bar{1}$ implies continuous exposure at all time points $t = 1 \dots 15$ at levels higher than the median exposure of 1.77 $\frac{mg}{m^3}$ while preventing leaving work when younger than 55. $\bar{a} = \bar{0}$ implies continuous exposure to $PM_{2.5}$ below 1.77 $\frac{mg}{m^3}$ while preventing leaving work when younger than 55. Smoking status and BMI measurements were missing for 1,512 (38%) of the 5,426 workers. $\psi_{n,1}$ is traditional IPCW-TMLE; $\psi_{n,2}$ is a multiple imputation type estimator; $\psi_{n,3}$ is a basic augmented IPCW-TMLE; $\psi_{n,4}$ is an augmented IPCW-TMLE with an iterative update of $Q_{X,n}$; $\psi_{n,5}$ is an augmented IPCW-TMLE where γ_n was estimated from multiply imputed data sets.

Chapter 4

Occupational Exposure to $PM_{2.5}$ and Incidence of Ischemic Heart Disease

4.1 Introduction

Particulate matter with an aerodynamic diameter of less than 2.5 μm ($PM_{2.5}$) is recognized as a major contributing factor to the global burden of heart disease, with the strongest evidence for cigarette smoke and air pollution sources. There are fewer studies of cardiovascular health and $PM_{2.5}$ at concentrations in the mid-range between active smoking and air pollution (Pope III, 2002). Studies of occupational exposures may help characterize the shape of the exposure-response curve, although $PM_{2.5}$ exposures vary widely across industry in terms of composition and temporal patterns of exposure. Moreover, the populations exposed to occupational $PM_{2.5}$ differ from those exposed to more general sources in terms of underlying health status, age, and other factors that may modify the health effects. Most occupational studies do not have information on important potential confounders, such as smoking and BMI, let alone measures of underlying cardiovascular health (Fang et al., 2010).

To address these research gaps, we have studied heart disease in a large cohort of actively employed aluminum production workers with extensive data on health status available from company personnel records, medical claims databases, and occupational medical records. A previous study (Costello et al., 2014) demonstrated a positive association between IHD incidence and current exposure to $PM_{2.5}$ in this cohort, but a protective effect of cumulative exposure. With extended follow-up, we have further investigated the relationship between cumulative exposure to $PM_{2.5}$ and IHD by applying a novel method to account for time-varying confounders on the causal pathway. Variables on the causal pathway between exposure at an earlier time period and heart disease can serve as confounders of the effect of current exposure on risk of heart disease (Greenland et al., 1999b). We believe the protective

association previously reported between cumulative exposure and IHD risk in (Costello et al., 2014) may be due, in part, to this phenomenon, an aspect of the healthy worker survivor effect (HWSE) (Eisen et al., 2006, Arrighi and Hertz-Picciotto, 1994). Workers with better health tend to accrue more exposure, through the preferential movement of workers with worse health to both lower exposed jobs as well as out of the work force. If poorer health was caused by prior exposure, standard statistical methods will not be able to generate consistent estimates of the effect of cumulative exposure (Naimi et al., 2011).

Robins and colleagues have developed a number of methods, known as the G methods, that can generate unbiased estimates in the presence of time-varying confounders on the causal pathway. The original work (Robins, 1986) was motivated by the problems of time-varying confounding in occupational health studies, although the methods have rarely been applied to occupational epidemiology. In recent years, several authors have used these methodologies, such as the parametric G formula (Cole et al., 2013), inverse probability weighting of marginal structural models (Dumas et al., 2013b), and G-estimation of accelerated failure time models (Chevrier et al., 2012b), to address this problem in occupational studies.

In the current work, we apply longitudinal TMLE to estimate the IHD incidence under hypothetical interventions to set cumulative exposure, adjusting for time varying confounding on the causal pathway. TMLE estimators (van der Laan and Rose, 2011, van der Laan and Rubin, 2006b) are semi-parametric efficient substitution estimators that use targeted fits of likelihood components. They also have the double robustness property, in that they remain unbiased if either of two likelihood components (the outcome models or the treatment models) are correctly specified. To our knowledge, this paper represents the first published application of longitudinal TMLE, or any doubly-robust method, to the field of occupational and environmental epidemiology.

4.2 Data

4.2.1 The Study Population and Outcome

Hourly workers employed at one of 11 US aluminum smelters and fabrication facilities for more than two years between 1/1/1996 and 12/31/2012 who were also enrolled in the company health plan were eligible for inclusion in the analysis. Before 2003, we assumed that all employed workers were enrolled in the company health plan because 97% of them filed a claim during this period. After 2003, when the company changed providers, active worker rolls were checked against an eligbility roster to determine health plan enrollment. Eligible workers were followed for incidence of IHD after a two-year washout period, implemented to remove prevalent cases of heart disease from the cohort. Follow-up ended at termination of

employment.

Workers were assigned to the smelter or fabrication sub-cohorts based on jobs held during follow-up. If they had ever been assigned to a smelter job they were included in the smelter sub-cohort and likewise for fabrication. Incident IHD was defined by any of the following events: i) insurance billing claim for a relevant procedure, such as revascularization, angioplasty, or a bypass, ii) face-to-face visit with a provider with a relevant ICD diagnosis code (410 - 414), iii) hospitalization for more than two days along with the relevant ICD admitting code, or iv) matching record of death from the National Death Index with the ICD-9 codes 410-414 or ICD-10 codes I-20 to I-25 listed in the cause of death field.

4.2.2 Exposure Assessment

The details of the exposure assessment have been previously described in Noth et al. (2013). In brief, each job was associated with an exposure level to total particulate matter (TPM) based off 8385 personal samples collected at 11 facilities between 1980 and 2011. Within eight of the facilities, additional samples were taken to determine the % $PM_{2.5}$ in the TPM. The % $PM_{2.5}$ was then multiplied by the TPM estimate to determine the mean concentration of $PM_{2.5}$ associated with a particular job. Additional modelling and expert judgment were used to generate estimates of TPM and % $PM_{2.5}$ from jobs without measured values. Each job was assigned a confidence level reflecting the method used to determine the exposure level.

Each worker's assigned exposure for a given year was the exposure level associated with the job they held on January 1st of that year. The current analysis was performed on subjects who, during any of their years of follow-up, had one of the two highest confidence levels for their exposure. This indicates that the TPM estimate that determined the $PM_{2.5}$ exposure concentration was based upon an actual measurement (i.e. not modeled). Exposure was treated as a binary variable in the analysis, each defined by a cut-off at either the median or 10th percentile exposure within each subcohort.

4.2.3 Covariates

Human resource records were the source for worker's age, sex, race, facility location, time since hire, job title, and job grade. Claims files from the primary health care provider were used to identify dates of diagnosis for four conditions associated with cardiovascular risk: diabetes, hypertension, dyslipidemia, and obesity. Claims files were also parsed by a proprietary algorithm (Verisk Health Inc, D_xCG Software) to compute a "risk score". The risk score estimated an individual's future likelihood of using medical services and served as a time-varying measure of overall worker health. This continuous measure was converted into deciles for the analysis. The risk score has been shown in prior research to predict a variety

of health outcomes including mortality in the higher deciles (Modrek and Cullen, 2013, Handel, 2011, Kubo et al., 2013, Modrek and Cullen, 2012). Smoking and BMI information was collected at occupational medicine clinics on site at each location.

4.3 Methods

Longitudinal TMLE allows for the estimation of cumulative incidence of disease in each year t in a cohort following a treatment regimen specified by the author (van der Laan and Gruber, 2011). We used a dichotomous definition of exposure in which $PM_{2.5}$ levels above a cut-off were defined as 'exposed', while $PM_{2.5}$ levels below the cut-off were defined as 'unexposed'. A priori, we chose two cut-offs which we calculate separately in each subcohort, one at the median exposure and one at the 10th percentile. We estimated the effect of remaining at work and in the same $PM_{2.5}$ exposure category throughout follow-up until retirement age, on the experience of incident IHD in the cohort.

We compare the estimated cumulative incidence of IHD within the worker population if they were all exposed above the cut-off during each year of follow-up to the estimated cumulative incidence within the same population if always exposed below the cut-off. Both treatment regimens include an intervention that prevents censoring. We define censoring in this population as leaving work prior to normal retirement age, or younger than 55 years old. We chose this treatment regimen (for both the exposure and the censoring mechanisms) in order to represent a realistic intervention on our population (van der Laan and Petersen, 2007, Bembom and van der Laan, 2007). This ensures the applicability of our results to workers under study and avoids sparse data problems due to inadequate numbers of workers following the regimens of interest.

4.3.1 Observed Data and Likelihood

Each observed worker history can be written as $O = (\bar{A}, \bar{L})$, where the overbar represents the history of a random variable, so $\bar{L} = (L(0), L(1), \ldots L(15))$ and $L(1)$ is L measured at the end of the first year of follow-up. $A(k) = (C(k), E(k))$ is the treatment node and contains $E(k)$, an indicator of exposure to $PM_{2.5}$ above a cut-off level during time point k, and $C(k)$, an indicator of remaining free from censoring, which is defined as leaving work when younger than age 55 during time point k. $L(k)$ contains all other variables, time variant and invariant, used in the analysis as measured at the end of time point k. $L(k)$ contains $Y(k)$, an indicator that a subject has been diagnosed with IHD prior to the end of time point k.

The likelihood of a data point can be written as:

$$\begin{aligned} P_O(O) &= \prod_{k=1}^{15} P_0(L(0))P_0(A(k)|\bar{L}(k-1),\bar{A}(k-1))P_0(\bar{L}(k)|\bar{A}(k)\bar{L}(k-1)) \\ &= \prod_{k=0}^{15} P_0(L(k)|\bar{L}(k-1),\bar{A}(k)) \prod_{k=1}^{15} P_0(A(k)|\bar{A}(k-1),\bar{L}(k-1)) \\ &= \prod_{k=0}^{15} Q_{0,L(k)}(L(k)) \prod_{k=1}^{15} g_{0,A(k)}(A(k)) \end{aligned}$$

where P_O is the distribution of O, $Q_{0,L(k)}$ is the true conditional distribution of $L(k)$ given $(\bar{L}(k-1),\bar{A}(k))$ and $g_{0,A(k)} = g_{0,E(k)}g_{0,C(k)}$ is the true conditional distribution of the treatment vector $(E(k),C(k))$ given $(\bar{L}(k-1),\bar{A}(k-1))$. We also use the notation $g_{0:k} = \prod_{j=0}^{k} g_{A(j)}$. The subscript 0 indicates that we refer to the true value of an object, and we use the subscript n to indicate that an object is an estimate based upon the observed data.

We define a statistical model, $\mathcal{M}$, for our observed data distribution P_0. $\mathcal{M}$ contains both $\mathcal{Q}$ the set of all possible values for $Q_0 = (Q_{0,L(0)}, \ldots Q_{0,L(15)})$ and $\mathcal{G}$,the set of all possible values of $g_0 = (g_{0,A(0)}, \ldots g_{0,A(15)})$. Therefore

$$P_0 \in \mathcal{M} = \{Q, g : Q \in \mathcal{Q}, g \in \mathcal{G}\}$$

with $\mathcal{Q}$ and $\mathcal{G}$ as defined above.

4.3.2 Target Parameter and Identifiability

A causal model serves as the link between the observed data and counterfactual data that would result from an intervention on the data generating system. We define our causal model using non-parametric structural equation models (Pearl, 1995). Let

$$L(k) = f_{L(k)}(\bar{L}(k-1),\bar{A}(k),U_{L(k)})$$

and

$$A(k) = f_{A(k)}(\bar{A}(k-1),\bar{L}(k-1),U_{A(k)})$$

where $f_{L(k)}$ and $f_{A(k)}$ are deterministic, non-parametric functions and the U elements represent the unobserved information used by nature to assign $L(k)$ and $A(k)$. We denote a counterfactual variable with a subscript a, as in Y_a, which is the random variable that would result from this system had exposure been set to $\bar{A} = \bar{a}$.

Our target statistical parameter is, $\Psi(P_0) = E_{P^a}(Y^a(t))$ where

$$P^a(l) = \prod_{k=0}^{t+1} Q_{0,L(k)}(l(k)|\bar{l}(k-1), \bar{A}(k-1) = \bar{a}(k-1))$$

represents the distribution of the observed data had we set the levels of $\bar{A} = \bar{a}$, i.e. set $C(k) = 0$ and $E(k) = e\ \forall k$. This is the G-computation formula for the post-intervention distribution. $L^a = (L(0), L^a(1), \ldots L^a(t))$ is a random variable distributed as P^a and $Y^a(t) \in L^a(t)$ is the outcome of interest measured at time t. Under a set of causal assumptions the distribution of variables distributed as $P^a(l)$ is equal to the distribution of the counterfactual variables $Y_a \sim P_a$ (Robins et al., 2000):

SRA : $L_a(t) \perp A(k)\big|\bar{L}(k), \bar{A}(k-1)\ \forall k < t$

Positivity : $P(A(k) = a\big|\bar{L}(k), \bar{a}(k-1)) > 0\ \forall\ (a, \bar{l}(k), \bar{a}(k-1))$

Consistency : $L_A = L$

The consistency assumption is implied by our use of non-parametric structural equation models (Pearl, 1995), but we include for completeness and because of its problematic nature in our application. We used a dichotomous exposure in this analysis, where $E = 1$ indicates an occupational exposure above a cut-off and $E = 0$ indicates exposure below. These definitions encompass a number of possible exposure values, and are an example of a compound treatment (Hernán and VanderWeele, 2011). Acknowledging this fact, a stronger consistency assumption must be made. This is treatment variation irrelevance, or that the counterfactual outcome for each subject would be the same if they were exposed at any level within a treatment definition. The relevance of the causal model to our observed data depends on this somewhat dubious assumption, but acknowledging this, we nonetheless believe that the statistical parameters we estimate are informative for our primary scientific question.

As first demonstrated by Robins, (Robins, 2000a), the mean outcome at time point t under intervention $\bar{A} = \bar{a}$ can be identified as a series of conditional expectations, the first of which takes the form:

$$\bar{Q}^a_{L(t)}(O) = E_0(Y(t)|\bar{L}(t-1), \bar{A}(t) = \bar{a}(t)).$$

This object corresponds to the regression of $Y(t)$ on the past covariates, performed among the population of treatment regimen followers (i.e. workers with observed $\bar{A}(t) = \bar{a}(t)$). This quantity, $\bar{Q}^a_{L(t)}$, can be sequentially regressed on the past in reverse chronological order, i.e. on $\bar{L}(k)$ for $k = (t-1, t-2, \ldots 0)$, amongst workers with observed $\bar{A}(k) = \bar{a}(k)$. We denote

this regression with

$$\bar{Q}^a_{L(k)} \equiv E_{Q_{L(k)}}(\bar{Q}^a_{L(k+1)}|\bar{L}(k-1), \bar{A}(k-1) = \bar{a}(k-1))$$

When $k = 1$ the result is a final constant $\bar{Q}^a_{L(1)}(O)$. Under the stated assumptions, we have that the distribution of the counterfactual outcome Y_a is equal to the distribution of the observed outcome under intervention, which equal to this final constant, a function of the observed data likelihood, or

$$E_0(Y_a(t)) = E_{P^a}(Y^a) = E\bar{Q}^a_{L(1)} = \Psi(P_0).$$

4.3.3 Efficient Influence Curve

For a given target parameter, $\Psi(P_0)$ an estimator is asymptotically efficient if and only if the estimator is asymptotically linear with influence curve equal to the canonical gradient $D^*(P_0)$ of the pathwise derivative of Ψ at P_0 in the model $\mathcal{M}$ (van der Laan and Rose, 2011). This gradient thus serves as a crucial building block for the construction of efficient estimators in general and TMLE in particular. One way to ensure that this property holds is for estimators to solve the efficient influence curve equation: $P_n D^*(Q_n, \psi_n) = 0$, where (Q_n, ψ_n) are estimates of likelihood components and the target parameter, respectively, and P_n represents the empirical distribution that places mass $\frac{1}{n}$ at each observed point O_i.

As first established by Bang and Robins (Bang and Robins, 2005) and expanded on in Van der Laan and Gruber (van der Laan and Gruber, 2012), the efficient influence curve for the mean outcome at time point t under intervention a can be written as $D^* = \sum_{k=0}^{t+1} D^*_k$. Here we have that

$$\begin{aligned} D^*_{t+1} &= \frac{I(\bar{A}(t) = \bar{a}(t))}{g_{0:t}}(Y - \bar{Q}^a_{L(t+1)}) \\ \text{and} & \\ D^*_k &= \frac{I(\bar{A}(k-1) = \bar{a}(k-1))}{g_{0:k-1}}(\bar{Q}^a_{L(k+1)} - \bar{Q}^a_k) \\ D^*_1 &= (\bar{Q}^a_{L(2)} - \bar{Q}^a_{L(1)}) \end{aligned}$$

Given consistently estimated conditional regressions $\bar{Q}^a_n \equiv (\bar{Q}^a_{L(t)}, \ldots \bar{Q}^a_{L(0)})$ and a consistent estimator g_n of the exposure and censoring mechanisms, we have that we can ensure the efficiency of the resulting estimator if

$$P_n \sum_{k=0}^{K+1} D^*_k(Q_n, g_n) = 0$$

4.3.4 The TMLE Algorithm

Implementation of TMLE for a this target parameter can begin once we have defined it and identified its efficient influence curve in the statistical model. The first step is the construction of initial estimators of the relevant likelihood components $\bar{Q}_0^a$ and g_0. TMLE then proceeds by defining a loss function $\mathcal{L}(\bar{Q}^a)$ for $\bar{Q}_0^a$ and a parametric submodel $\{\bar{Q}^a(\epsilon, g) : \epsilon\}$ so that the linear span of the score of the loss function applied to the submodel $\frac{\partial}{\partial \epsilon}\mathcal{L}(\bar{Q}^a(\epsilon, g))$ at $\epsilon = 0$ includes the efficient influence curve $D^*(Q, g)$. There is a one:one correspondence between the components of the conditional regressions, $\bar{Q}^a$ and components of the efficient influence curve. This allows for the creation of $k = 1, \ldots, t$ loss functions for each component of $\bar{Q}^a$, $\mathcal{L}_k(\bar{Q}^a_{L(k)})$, as well as t corresponding submodels $\bar{Q}^a_{L(k)}(\epsilon, g)$ such that $\frac{\partial}{\partial \epsilon}\mathcal{L}_k(\bar{Q}^a_{L(k)}(\epsilon, g))$ at $\epsilon = 0$ equals the kth component D^*_k of the efficient influence curve D^*.

These loss functions and submodels are used to estimate and update each component of $\bar{Q}^a$ in the following manner. First, an initial estimator $\bar{Q}^a_{L(t),n}$ of $\bar{Q}^a_{L(t),0}$ is made by regressing Y on $\bar{L}(k), \bar{A}(k) = \bar{a}(k)$. This initial estimator is then updated by minimizing the empirical mean of the tth loss function $\mathcal{L}_t(\bar{Q}^a_{L(t)})$, resulting in an updated fit $\bar{Q}^{a,*}_{L(t),n} = \bar{Q}^a_{L(t),n}(\epsilon_{t,n}, g_n)$ where $\epsilon_{t,n} = \text{argmin}_\epsilon P_n \mathcal{L}_t(\bar{Q}^a_{L(t),n}(\epsilon, g_n)(O))$. An initial estimator for the next member of $\bar{Q}^a$, $\bar{Q}^a_{L(t-1),n}$ is then generated by regressing $\bar{Q}^{a,*}_{L(t),n}$ on $\bar{L}(t-1), \bar{A}(t-1) = \bar{a}(t-1)$. The updating process is then repeated to generate $\bar{Q}^{a,*}_{L(t-1),n} = \bar{Q}^a_{L(t-1),n}(\epsilon_{t-1,n}, g_n)$ where $\epsilon_{t-1,n} = \text{argmin}_\epsilon P_n \mathcal{L}_{t-1}(\bar{Q}^a_{L(t-1),n}(\epsilon, g_n)(O))$. This continues until an initial estimate for $\bar{Q}^a_{L(1),n}$ is created and updated as $\bar{Q}^{a,*}_{L(1),n} = \bar{Q}^a_{L(1),n}(\epsilon_{0,n}, g_n)$ where $\epsilon_{1,n} = \text{argmin}_\epsilon P_n \mathcal{L}_1(\bar{Q}^a_{L(1),n}(\epsilon, g_n)(O))$. $n^{-1}\sum_{i=1}^n \bar{Q}^{a,*}_{L(1),n}(O)$ is the final TMLE of the target parameter $\Psi(\bar{Q}^a_0) = E_{P_0}(Y^a)$. The minimization of the empirical mean of the loss functions ensures that $Q^a_{L(k),n}, k = 1, \ldots, t$ solves the score equations for each of the submodels, and therefore we have that $(\bar{Q}^{a,*}_n, g_n)$ solves the efficient influence curve equation:

$$P_n D^*(\bar{Q}^{a,*}_n, g_n)(O) = 0.$$

4.3.5 Practical Implementation

The following steps detail how the implementation of TMLE in our application for a given treatment regimen, $\bar{A} = \bar{a}$ and time point t.

- We first generated estimators for the treatment mechanism $g_n = (g_{E,n}, g_{C,n})$, containing exposure assignment and censoring mechanisms. $g_{E,n}$ was estimated using main term logistic regression and regressing $E(t)$ on $L(t-1), \bar{A}(t-1) = \bar{a}(t-1), t$ among all active workers and time periods t. $g_{C,n}$ was estimated using main term logistic regression to regress $C(t)$ on $L(t-1), E(t), \bar{A}(t-1) = \bar{a}(t-1), t$ among all active workers younger than 55 years old and time periods t. For both fits, we made a Markov assumption

that the values of the time-varying confounders as measured at the most recent time points were the only salient values.

- We then regressed $Y(t)$ onto $\bar{A}(t) = \bar{a}(t), L(t-1)$, using either main term logistic regression or alternatively bi-directional stepwise regression if the total number of workers with $Y(t) = 1$ was less than 250, giving us the regression object $\bar{Q}^a_{Y,n}$

- We then fluctuated this initial estimator to target the parameter of interest. We used the initial estimator $\bar{Q}^a_{Y,n}(O)$ as an offset in a univariate logistic regression of $Y(t)$ on $I(\bar{A}(t) = \bar{a}(t))/g_{0:t,n}$ among the subjects with $\bar{A}(t) = \bar{a}(t)$. This regression object is the TMLE $\bar{Q}^{a,*}_{Y,n}(O)$ of the first conditional regression, $\bar{Q}^a_{Y,0}(O)$.

- Next, we ran a logistic regression of $\bar{Q}^{a,*}_{Y,n}(O)$ onto $\bar{A}(t-1) = \bar{a}(t-1), L(t-1)$, again choosing between main term logistic regression and bi-directional stepwise regression, resulting in $\bar{Q}^a_{L(t-1),n}(O)$.

- We fluctuated this estimator, $\bar{Q}^a_{L(t-1),n}(O)$, by using it as an offset in the regression of $\bar{Q}^{a,*}_{Y,n}$ on $I(\bar{A}(t-1) = \bar{a}(t-1))/g_{0:t-1,n}$. The result of this is the TMLE $\bar{Q}^{a,*}_{L(t-1),n}$ of $\bar{Q}^{a,*}_{L(t-1),0}$.

- This process continued for $k = (t-2, \ldots, 1)$ wherein we ran a logistic regression of the previous TMLE fit $\bar{Q}^{a,*}_{L(k+1),n}(O)$ onto $\bar{A}(k) = \bar{a}(k), L(k)$ among the population of treatment regimen followers at time point k. We then updated this fit by using the initial estimate, $\bar{Q}^a_{L(k),n}(O)$ as an offset of the regression of $\bar{Q}^{a,*}_{L(k+1),n}(O)$ on $I(\bar{A}(k) = \bar{a}(k))/g_{0:k,n}(O)$, resulting in the TMLE, $\bar{Q}^{a,*}_{L(k),n}$, of that conditional regression $\bar{Q}^a_{L(k),0}$.

- The final step left us with $\bar{Q}^{a,*}_{L(1),n}(O)$, which was a function of $L(0)$ only and a series of nested conditional regressions $Q^*_n = (Q^*_{L(t),n}, \ldots, Q^*_{L(1),n})$. We estimate $\bar{Q}^a_{L(0}$ by taking the average of this function applied to the entire worker population: $\frac{1}{n}\sum_{i=1}^n \bar{Q}^{a,*}_{L(1),n}(L_i(0)) = \bar{Q}^{a,*}_{L(0),n} \equiv \psi_n$ is the TMLE of our target parameter $\Psi(\bar{Q}^a_0)$.

- We use influence curve based variance estimates for inference, so $\hat{\sigma}^2_{IC} = \hat{\sigma}^2_n/n$, where

$$\hat{\sigma}^2_n = \mathrm{Var}(D^*(Q_n, g_n, \psi_n)(O_i))$$

This series of iterative regressions was performed separately for each time period ($t = 1, \ldots 15$) to create estimates of the cumulative incidence of disease among the whole population at time t. These 15 estimates were then used to create marginal incidence curves which estimate the experience of the cohort over the entire length of follow-up under the specified treatment regimen. We also used these estimates to calculate average treatment effects and rate ratios and their corresponding confidence intervals, comparing the regimens with exposures over the cut-off to the regimens with exposures under the cut-off. For each of the four estimation procedures (two subcohorts each with two binary exposure variables),

fits of the g models ($g_{A,n}$ and $g_{C,n}$) were generated using all regimen-following person-years. These model fits were then used to perform the outcome model updates for each of the 15 time points. The analysis was performed using a modified version of the `ltmle` (Schwab et al., 2013, Lendle et al., 2014) package in R (R Core Team, 2013b) version 3.0.2.

For higher values of t, there were few subjects still at risk, which created the potential for overfitting within the outcome model. We used bi-directional stepwise regression by AIC to perform variable selection for outcome models with fewer than 250 cases (Peduzzi et al., 1996) while forcing risk score and current exposure into the model. This allowed our models to be more parsimonious and ensured that there was still some error variance which the targeting step needs to be effective. We note that these model choices were made in order to reduce the computational complexity and increase model interpretability, and recognize that they open up the potential for bias due to model misspecification.

4.3.6 Incorporation of Multiple Imputation

Some covariates were missing in a portion of the cohort population, and we used multiple imputation to account for this. Multiple imputation involves the combination of results from different data sets in which the missing covariates are filled in by different imputation models. Rubin's rules (Rubin, 1987) can be used to combine these estimates as long as the estimator of interest has an asymptotically normal distribution. As demonstrated by Van der Laan and Gruber (van der Laan and Gruber, 2012),

$$(\psi_n - \psi_0) = \frac{1}{n}\sum_{i=1}^{n} D^*(Q_n, g_n, \psi_n)(O_i) + o_p\left(\frac{1}{\sqrt{(}n)}\right),$$

which implies, given our model, that the TMLE is asymptotically normal

$$\sqrt{(}n)(\psi_n - \psi_0) \to \mathcal{N}(0, \sigma^2_{IC}).$$

We described above the process for implementing TMLE for a single data set, resulting in an estimator ψ_n and variance σ^2_n. Given data sets and estimators indexed by $b = 1 \ldots B$, where $B = 5$ we combined the estimates from each as follows.

$$\psi_n = \frac{1}{B}\sum_{b=1}^{B} \psi_{n,b}$$

was the reported estimate and

$$\sigma^2_{IC} = \frac{1}{B}\sum_{b=1}^{B}\sigma^2_{IC,b} + \frac{1}{B-1}\sum_{b=1}^{B}\left(\psi_{n,b} - \psi_n\right)^2$$

was the variance used for inference.

Complete information was not available on all of the covariates of interest for all workers. Smoking status was missing for 51% of person-years, BMI (body mass index) for 23%, marital status for 2%, and risk score for 15% of person-years. We did not apply the augmented IPCW-TMLE procedure we discussed in chapter 3 because the patterns of missing data were not monotonic, in that a missing value for any of the variables subject to missingness did not predict the missingness of any of the other variables. We could have used a hybrid approach, in which the imputation was used for some variables, or to create a monotone missingness pattern, and augmented IPCW-TMLE was used for some of the other variables. This is what we did in the applied example presented in section 3.6 of this dissertation. Multiple imputation was performed using the `proc mi` procedure in SAS 9.3 (SAS Institute Inc., 2011), and included all variables used in our analysis in the prediction model.

4.4 Results

The original cohort contained 16,991 workers (140,179 person-years) and the restriction to only workers with a high-confidence exposure resulted in an analysis cohort of 13,529 workers and 112,293 person-years, roughly 80% of the original cohort. The smelter subcohort included 5,527 workers (46,723 person-years) and the fabrication subcohort included 7,211 workers (61,375 person-years). Some workers worked only in other environments, such as refineries or mines, and were not included in either subcohort while 680 workers worked in both smelters and fabricators and were included in both. In the smelter sub-cohort, the median $PM_{2.5}$ concentration was 1.77 $\frac{mg}{m^3}$ and the 10th percentile was 0.16 $\frac{mg}{m^3}$. For the fabrication subcohort, the median $PM_{2.5}$ concentration was 0.20 $\frac{mg}{m^3}$ and the 10th percentile was 0.06 $\frac{mg}{m^3}$.

We compare the baseline demographic characteristics of cohort members by facility type and exposure categories defined by alternative cut-offs in tables 4.1 and 4.2. In both smelters and fabrication facilities, workers exposed above the median cut-off have lower frequencies of cardiovascular risk factors and other measures of overall health than workers exposed below the cut-off, although the rates of IHD are similar or slightly higher among the exposed. At the 10th percentile cut-off, these differences become more stark. These tables are consistent with a pattern of workers tending to move to lower exposed jobs as cardiovascular risk, job tenure and age increase.

Table 4.3 presents the smelter and fabrication worker population sizes and incident disease

counts by year of follow-up, with and without the restriction to those following the treatment regimen of staying in the same exposure category (for the median cut-off). This restriction resulted in the loss of 14% of the person-years and 12% of the incident cases from the fabrication analysis and in the loss of 17% of the person-years and 16% of the incident cases from the smelter analysis.

Table 4.4 shows the model parameters from the logistic regressions used to estimate the effect at the median cut-off in the smelter sub-cohort. It contains parameter values for the treatment and censoring models $g_{A,n}, g_{C,n}$, as well as two of the outcome models for the $t = 15$ estimator. The parameter signs are generally consistent with the hypothesis of the healthy worker survivor effect operating in this population. The results from the 10th percentile cut-off in the smelter sub-cohort and both cut-offs for the fabrication facility sub-cohort (not shown) provided similar evidence of the direction of effects.

Figures 4.1, 4.2, 4.3, and 4.4 contain the marginal cumulative incidence curves as estimated with TMLE for each of the four analysis groups. Each curve estimates the percentage of the cohort that would remain undiagnosed with heart disease by the end of follow up had all workers followed the treatment regimen. These curves are not traditional survival curves, in that they do not purport to estimate the cohort's absolute freedom from IHD incidence. Rather they estimate the cohort's freedom from observed IHD, specifically IHD incidence prior to leaving work when older than 55.

Table 4.5 contains the average treatment effects (ATE) and causal rate ratios (RR) at year 15 for each of the four analysis groups. The ATE is the difference between the cumulative incidence of ischemic heart disease as predicted for a cohort subjected to the treatment regimen with exposure above the cut-off and the cumulative incidence for that same cohort subjected to the treatment regimen exposed below the cut-off. At year 15, we estimate that the smelter worker sub-cohort, if constantly exposed above the median cut-off of 1.77 $\frac{mg}{m^3}$ while remaining at work until 55, would experience a 2.1% (95% CI = (-1.3%, 5.5%)) higher incidence of IHD compared to the same cohort if constantly exposed below the cut-off. For the 10th percentile cut-off of 0.16 $\frac{mg}{m^3}$ in the smelter sub-cohort, we estimate that the cumulative incidence of IHD would be higher by 2.9% (0.6%, 5.1%). Among the fabrication cohort, we estimate an ATE of 0.9% (-1.6%, 4.1%) for the median cut-off of 0.20 $\frac{mg}{m^3}$ and an ATE of 2.5% (0.8%, 4.1%) for the 10th percentile cut-off of 0.06 $\frac{mg}{m^3}$.

The estimation of marginal cumulative incidence also allows us to calculate causal risk ratios for the same comparison groups, by taking the ratio of the two incidences. Among the smelter sub-cohort, the average causal risk ratio (over the 15 years of follow-up) was 1.39 (0.81, 2.39) at the median cut-off and 1.77 (1.03, 3.06) at the 10th percentile cut-off. Among the fabrication facility sub-cohort, the average causal risk ratio was 1.14 (0.80, 1.63) at the median cut-off and 1.45 (1.13, 1.86) at the 10th percentile cut-off.

4.5 Discussion

These results provide evidence that increased risk of IHD is associated with occupational exposure to $PM_{2.5}$ in both the fabrication and smelter sub-cohorts. We were able to adjust for possible time-varying confounders on the causal pathway through our use of the longitudinal TMLE procedure. We have not addressed the question of precisely when the biologically relevant exposure occurred; i.e. if IHD risk during time t is more dependent on exposure at time t or on exposures accumulated prior to time t. This question would be answered by estimating the indirect effect of cumulative exposure that does not travel through current exposure, a target parameter that deserves future research.

The magnitude of the additive difference was similar in the smelters using either the median or 10th percentile cut-off. By contrast, we saw a larger difference for the 10th percentile compared to the median cut-off in fabrication. These results are consistent with a linear exposure -response curve over the higher concentrations of the smelters and a flattened exposure-response curve over the lower concentrations of fabrication facilities. One possible explanation is that the measured variables were better able to adjust for time-varying confounding within the smelter population and that the flattened exposure-response curve is a result of uncorrected bias due to HWSE in fabrication (Stayner et al., 2003). Pathway analysis (not shown) and anecdotal evidence suggests that the selection effects of the measured variables are stronger in the smelter environment because workers are screened and jobs with heat exposure are restricted to workers with low cardiovascular risk. Future research with this cohort will involve investigating these exposure response curves further through the estimation of the parameters of marginal structural models.

A sensitivity analysis demonstrated that including workers who never had a high-confidence exposure estimate reduced the effect estimates. The industrial hygiene measurements used to determine the job exposure matrix and the confidence scores were collected more frequently in areas where high exposures were expected. Thus, subjects with high exposures were preferentially selected into our final cohort, although many lower exposure jobs were still measured with high confidence. If workers in low confidence jobs were substantially different from the rest of the population then the restriction could result in biased effect estimates. It is also possible that the difference in effect estimates is due to reduced exposure misclassification in the restricted cohort.We believe that the second conditions is more likely because we are certain that the restriction reduced exposure misclassification, while there is no reason to believe that the confidence measure is associated with the disease process. Therefore we feel that the high-confidence restriction represents a less biased analysis than the full cohort.

We want to control for leaving work when it is possibly mediated by health status caused by prior exposure. Censoring is defined in this analysis as leaving work when younger than 55. Modeling (not shown) indicates that, at each age, workers who retired had worse health

predictors than those who did not. The choice of age 55 represents a compromise between controlling for health as a time-varying predictor of leaving work and reducing the reliance of the procedure on those unusual individuals who stay at work past their eligibility for a full retirement. Sensitivity analysis demonstrated that the results were robust to the age cut-off; changing the age to 60 or 62 did not substantially change the results. Continuing follow-up after work termination would give us the information needed to study the effects of $PM_{2.5}$ on post-retirement health.

We observed higher crude rates of IHD among the fabrication workers, and the marginal cumulative incidence estimates reflect this fact. As with any procedure, these estimates do not generalize precisely to different populations with their own underlying risk and termination patterns. For example, heart disease rates among the fabrication workers, had they been exposed to the $PM_{2.5}$ in the smelters instead, cannot be inferred. Although the two sub-cohorts exhibit similar rates of chronic diseases, such as diabetes and hypertension, the summary risk score was higher among the fabrication workers.

We observed excess IHD risk associated with $PM_{2.5}$ in both smelters and fabrication facilities where the composition and particle size distribution differ. In fabrication, the $PM_{2.5}$ is composed mostly of water-based metalworking fluids and in smelters, of inorganic materials, such as fluorides, alumina dust, metals and related fumes (Ronneberg, 1995, Noth et al., 2013). Thus our findings suggest that the total mass of $PM_{2.5}$ may be the common causal agent, although further study is needed to address this question. It is also possible that the observed relationship between exposure to $PM_{2.5}$ and IHD could be due in part to co-exposure to other known cardiovascular hazards in these workplaces, such as noise and heat.

The goal of causal inference is to make inference about counterfactual quantities, and unbiased estimation can proceed only if several assumptions are met (Robins et al., 2000). We believe there is minimal unmeasured confounding in this analysis. We have a rich data set that captures many of the salient aspects of health upon which workers might base their employment decisions. We also believe that we have limited positivity violations; there were no combinations of covariates that strongly predicted exposure status. We believe the assumption of consistency, which is often subsumed by the use of non-parametric structural equation models (Pearl, 1995), may be more problematic. The dichotomization of exposure means that a range of different true exposure values and constituencies are contained in a single category. It is probably not the case that all workers would have the same outcome had they been assigned to any exposure level above the cut-off that defined the category.

Longitudinal TMLE estimators have the property of double-robustness, in that they remain unbiased if either the outcome models or the treatment (exposure and censoring) models are correctly specified. For the sake of simplicity and computational efficiency, we chose to proceed with main-term logistic regression with limited use of variable selection. Our model fits could be improved and bias possibly reduced if we used, for instance, a cross-

validated ensemble learner (van der Laan et al., 2007). By implementing multiple imputation, we assumed missingness at random (Kenward and Carpenter, 2007) for the missing variables as well as a specific model form for the imputation model. Violations of this assumption or misspecification of the imputation model could result in bias. A sensitivity analysis showed that the results were robust to the removal of the smoking and BMI variables, indicating that they function as limited confounders in this data set.

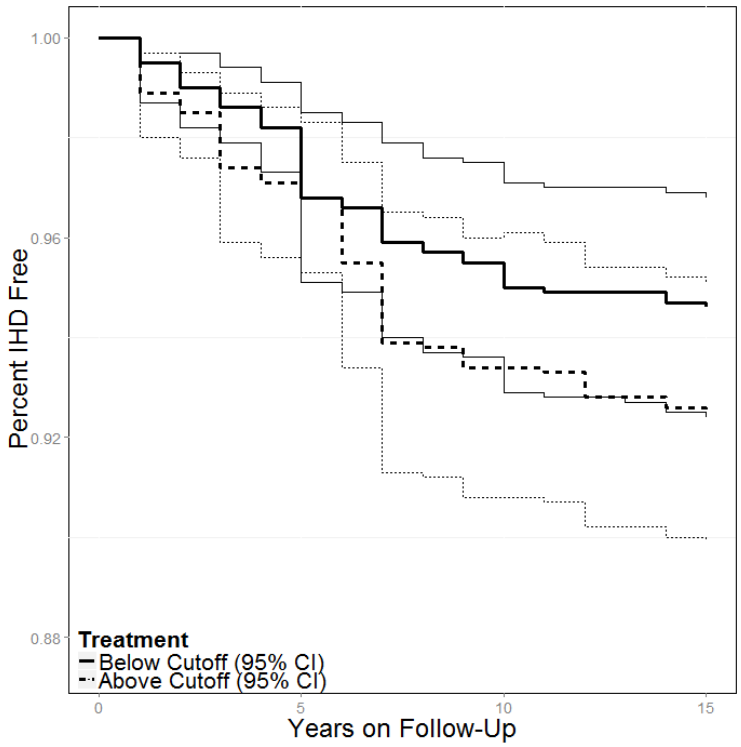

Figure 4.1: Estimated cumulative survival from IHD and 95% confidence iIntervals, adjusted for measured baseline and time-varying risk factors, among the smelter worker population if continuously exposed vs unexposed at the median cut-off of 1.77 $\frac{\text{mg}}{\text{m}^3}$

Table 4.1: Smelter worker cohort demographics, time varying covariates and outcomes by $PM_{2.5}$ exposure cut-off and exposure level at baseline

PM2.5 Cutoff	**Median (1.77 mg/m3)**		**10th %ile (0.16 mg/m3)**		**-**
Exposure Status at Baseline	**Above**	**Below**	**Above**	**Below**	**Total**
N	2,808	2,618	4,914	512	5,426
Person-years	21,241	18,772	35,827	4,186	40,013
Follow-Up Time, mean (IQR)	7.6 (4 - 11)	7.2 (3 - 11)	7.3 (3 - 11)	8.2 (4 - 12)	7.3 (3 - 11)
Demographics					
Male (%)	96%	93%	95%	92%	94%
White (%)	87%	84%	85%	91%	86%
Age, median (IQR)	42 (34 - 51)	45 (36 - 52)	43 (34 - 51)	47 (42 - 53)	44 (35 -51)
Ever Been Married (%) [1]	85%	84%	84%	85%	84%
Time Varying Covariates					
Time Since Hire, median (IQR)	8 (2 - 24)	13 (2 - 25)	8.5 (2 - 25)	22 (6 - 27)	13.5 (2 -25)
Proportion of year off work, mean	4%	4%	4%	3%	4%
High Job Grade (%)	36%	32%	33%	45%	34%
Hypertension (%)	13.5%	13.7%	13%	16%	14%
Diabetes (%)	4.0%	5.0%	4.4%	5.5%	4.5%
Dyslipidemia (%)	11%	14%	12%	16%	12%
Clinically Obese (%)	1.3%	1.0%	1.2%	0.6%	1.1%
Risk Score Decile, mean (IQR) [1]	4.6 (2 - 7)	5.0 (3 - 7)	4.7 (2 - 7)	5.7 (3 - 8)	4.8 (2 - 7)
BMI, median (IQR) [1]	28.3 (25 - 32)	28.9 (26 - 32)	28.6 (26 - 32)	28.4 (26 - 32)	28.6 (26 - 32)
Smoking Status: Current (%) [1]	25%	28%	27%	27%	27%
Smoking Status: Ever (%) [1]	38%	33%	35%	33%	35%
Smoking Status: Never (%) [1]	38%	39%	38%	40%	38%
Cumulative Exposure ($\frac{mg}{m^3}$*years), median (IQR)	19.2 (6 - 62)	5.1 (3 - 18)	12.6 (5 - 44)	3.4 (1 - 5)	27.2 (4 - 41)
Current Exposure ($\frac{mg}{m^3}$), median (IQR)	3.3 (2.0 - 3.4)	0.6 (0.3 - 1.5)	2.3 (1.1 - 2.6)	0.12 (.07 - .16)	2.1 (0.6 - 2.6)
Outcomes of Interest					
Incident IHD, n (%)	212 (7.5%)	191 (7.3%)	364 (7.4%)	39 (7.6%)	403 (7.4%)
Censored, n (%)	443 (15.7%)	439 (16.8%)	813 (16.5%)	69 (13.5%	882 (16.2%)

[1] Among workers with recorded values. Marital status was measured in 98% of smelter workers, risk score in 80%, BMI in 71%, and smoking status in 40%.

Table 4.2: Fabrication worker cohort demographics, time varying covariates and outcomes by $PM_{2.5}$ exposure cut-off and exposure level at baseline

PM2.5 Cutoff	**Median (0.19 mg/m3)**		**10th %ile (0.06 mg/m3)**		**-**
Exposure Status at Baseline	**Above**	**Below**	**Above**	**Below**	**Total**
N	3,344	3,777	6,612	509	7,121
Person-years	26,346	25,339	47,473	4,212	51,685
Follow-Up Time, mean (IQR)	7.9 (4 - 13)	6.7 (3 - 10)	7.2 (4 - 11)	8.3 (4 - 13)	7.3 (4 - 11)
Demographics					
Male (%)	88%	74%	81%	82%	81%
White (%)	80%	84%	82%	91%	82%
Age, median (IQR)	42 (35 - 51)	45 (38 - 53)	44 (36 - 52)	45 (39 - 52)	44 (36 - 52)
Ever Been Married (%) [1]	80%	74%	77%	76%	77%
Time Varying Covariates					
Time Since Hire, median (IQR)	8 (2 - 21)	13 (2 - 25)	9 (2 - 24)	14 (2-25)	9 (2 - 24)
Proportion of year off work, mean	3%	4%	3%	4%	3%
High Job Grade (%)	41%	43%	43%	32%	42%
Hypertension (%)	10.7%	11.9%	11%	16%	11%
Diabetes (%)	3.6%	5.1%	4.2%	6.3%	4.4%
Dyslipidemia (%)	10%	13%	11%	14%	12%
Clinically Obese (%)	0.9%	1.0%	0.9%	1.2%	0.9%
Risk Score Decile, mean (IQR) [1]	4.9 (2 - 7)	5.3 (3 - 8)	5.1 (3 - 8)	5.2 (3 - 7)	5.1 (2 - 7)
BMI, median (IQR) [1]	29.2 (26 - 33)	28.5 (25 - 32)	28.9 (26 - 33)	28.4 (26 - 33)	28.8 (26 - 33)
Smoking Status: Current (%) [1]	28%	28%	27%	34%	28%
Smoking Status: Ever (%) [1]	30%	30%	30%	28%	35%
Smoking Status: Never (%) [1]	42%	43%	43%	38%	42%
Cumulative Exposure (mg/m3*years), median (IQR)	3.9 (1 - 9)	2.9 (0.3 - 3)	2.3 (0.7 - 5)	2.2 (0.2 - 1)	2.0 (1 - 5)
Current Exposure (mg/m3), median (IQR)	0.89 (.25 - .97)	0.12 (.07 -0.14)	0.21 (.14 - .45)	0.04 (.04 - .06)	0.48 (.12 - .37)
Outcomes of Interest					
Incident IHD, n (%)	294 (8.8%)	279 (7.4%)	531 (8.0%)	42 (8.3%)	573 (8.0%)
Censored, n (%)	551 (16.5%)	548 (14.5%)	1025 (15.5%)	74 (14.5%)	1099 (15.4%)

[1] Among workers with recorded values. Marital status was measured in 95% of fabrication workers, risk score in 79%, BMI in 69%, and smoking status in 44%.

Table 4.3: Worker cohort membership and incident ischemic heart disease cases by year of follow-up and facility type for all workers and only workers exposed consistently to either above or below the median (fabricators: 0.19 mg/m3; smelters: 1.77 mg/m3) level of PM2.5

	Fabricators				**Smelters**			
	All Workers		Workers With Constant Exposure		All Workers		Workers With Constant Exposure	
Time On Follow Up	Subjects	Incident Cases	Subjects	Incident Cases	Subjects	Incident Cases	Subjects	Incident Cases
1	7121	80	7121	80	5426	39	5426	39
2	6623	75	6246	70	5044	56	4656	53
3	5824	63	5346	56	4498	47	3962	41
4	5380	52	4758	50	4047	35	3404	31
5	4685	57	4015	49	3639	45	2955	38
6	3716	35	3061	30	3153	35	2452	30
7	3193	42	2579	36	2731	31	2064	26
8	2757	28	2197	21	2417	23	1800	19
9	2536	27	1996	24	2185	23	1582	16
10	2289	22	1761	13	1912	23	1353	14
11	1916	13	1443	12	1563	13	1066	9
12	1774	20	1324	16	1173	10	792	8
13	1501	25	1090	21	927	9	633	5
14	1313	26	950	20	725	10	498	7
15	1057	8	763	6	573	4	396	4
Overall	**51685**	**573**	**44650**	**504**	**40013**	**403**	**33039**	**340**

Table 4.4: Model Parameters for logistic regression models estimated among smelter cohort using median cut-off (1.77 mg/m3). Treatment model and censoring model estimate probability of receiving high exposure and remaining uncensored, respectively. Outcome models predict probability of being an observed case prior to time point 15 given covariates measured at time t

		Treatment Model	Censoring Model	Outcome Model (t=1)	Outcome Model (t=15)
		Estimate	**Estimate**	**Estimate**	**Estimate**
	Intercept	-4.22^*	2.57^*	-4.50^*	1.54
	Male	0.16	-0.35^*	0.28^*	-
	White	0.27^*	-0.03	0.01	-
Facility Codes	ALC	-0.18	0.65	0.23	-
	FWX	-0.27	-0.57	0.34^*	-
	MAS	1.74^*	0.98	0.54^*	-
	ROK	2.10^*	-0.88	0.29	-
	WAR	0.39	0.85^*	0.20	-
	Ever been Married	-0.02	0.51^*	0.05	-
	Obesity Diagnosis (t)	0.31	1.30	-0.03	-
	Obesity Diagnosis (t - 1)	-0.74	-0.70	0.09	-
	Diabetes Diagnosis (t)	0.14	0.89	0.29^*	3.71^*
	Diabetes Diagnosis (t-1)	-0.15	-0.83	0.16	-
	Hypertension Diagnosis (t)	-0.20	0.13	-0.09^*	-
	Hypertension Diagnosis (t-1)	0.26	-0.34	0.21^*	-
	Dyslipidemia Diagnosis (t)	0.07	-0.05	0.46^*	-2.45^*
	Dyslipidemia Diagnosis (t-1)	-0.12	-0.01	0.06	-
	Risk Score Decile (t)	0.00	-0.07^*	0.01	0.42
	Risk Score Decile (t-1)	-0.04^*	-0.03^*	0.01	-
	BMI	0.00	0.00	0.00	-
	Smoking Status (Current vs Ever/Never)	-0.11	0.06	0.07^*	-
	High Job Grade	-0.27^*	-0.09	-0.03	-
	Calendar Year	0.05^*	-0.11^*	-0.09^*	-
	Age	0.02^*	0.03^*	0.03^*	-0.19^*
	Time Since Hire	0.00	0.03^*	0.00^*	-
	Time Since Follow Up Start	-0.01	0.01	-	-
	Exposure in Year t-1	6.54^*	0.08	0.10^*	-
	Cumulative Exposure	0.00^*	-0.01	0.00	-
	Exposure in Year t	-	0.00^*	0.05	0.84

[1] Outcome models at time t use stepwise AIC to select variables in the regression model if there were less than 250 subjects subjects diagnosed with IHD during times $(t \ldots, 15)$

[2] $*:\ p < 0.1$

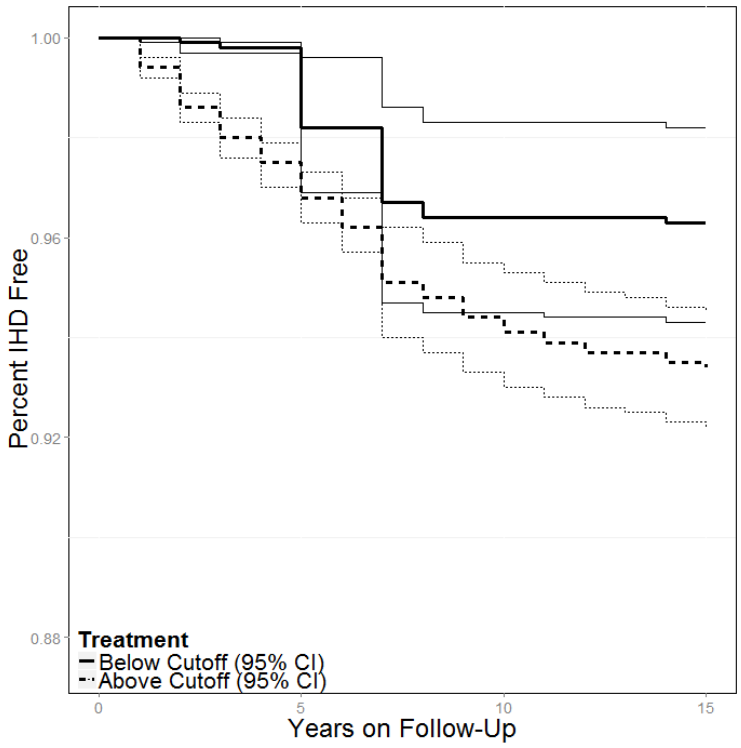

Figure 4.2: Estimated cumulative survival from IHD and 95% confidence intervals, adjusted for measured baseline and time-varying risk factors, among the smelter worker population if continuously exposed vs unexposed at the 10th percentile cut-off of 0.16 $\frac{\text{mg}}{\text{m}^3}$

Facility		**ATE**		**RR**	
Type	Cut-off	Estimate	95% CI	Estimate	95% CI
Smelter	Median (1.77 $\frac{mg}{m^3}$)	.021	(-.013, .055)	1.39	(0.81, 2.39)
Smelter	10^{th}(0.16 $\frac{mg}{m^3}$)	.029	(-.006, .051)	1.77	(1.03, 3.06)
Fabricator	Median (0.20 $\frac{mg}{m^3}$)	.009	(-.016, .035)	1.14	(0.80, 1.63)
Fabricator	10^{th} (0.06 $\frac{mg}{m^3}$)	.025	(.008, .041)	1.45	(1.13, 1.86)

Table 4.5: Average treatment effects (ATE) and Risk Ratios (RR) of occupational exposure to PM2.5 by facility type and cut-off level. The ATE is the difference between the cumulative incidence ischemic heart disease predicted for a cohort subject to continuous exposure above the cut-off and the incidence predicted for the same cohort subject to constant exposure below that cut-off, where in both cohorts workers work until retirement age. The RR is the ratio between the two cumulative incidences.

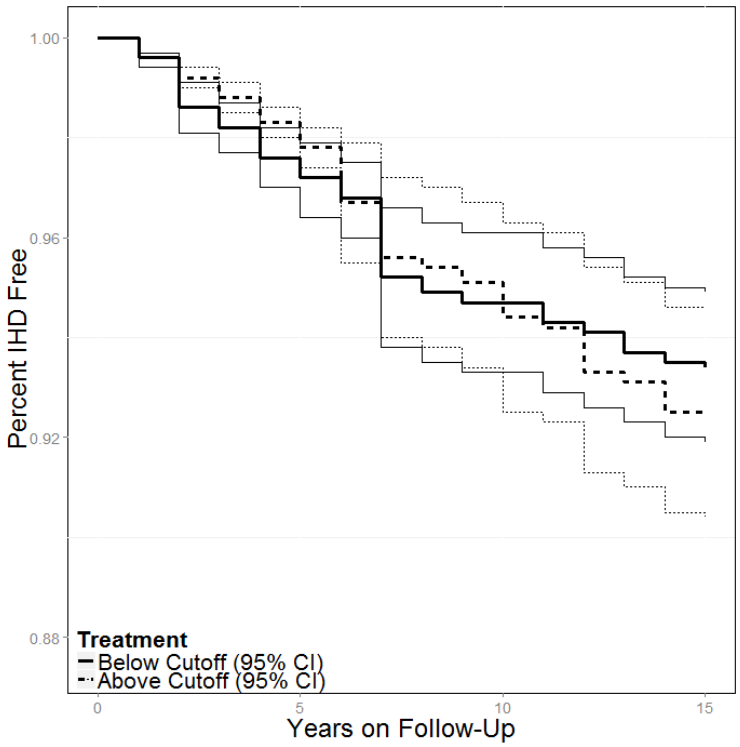

Figure 4.3: Estimated cumulative survival from IHD and 95% confidence intervals, adjusted for measured baseline and time-varying risk factors, among the fabricator worker population if continuously exposed vs unexposed at the median cut-off of 0.20 $\frac{\text{mg}}{\text{m}^3}$

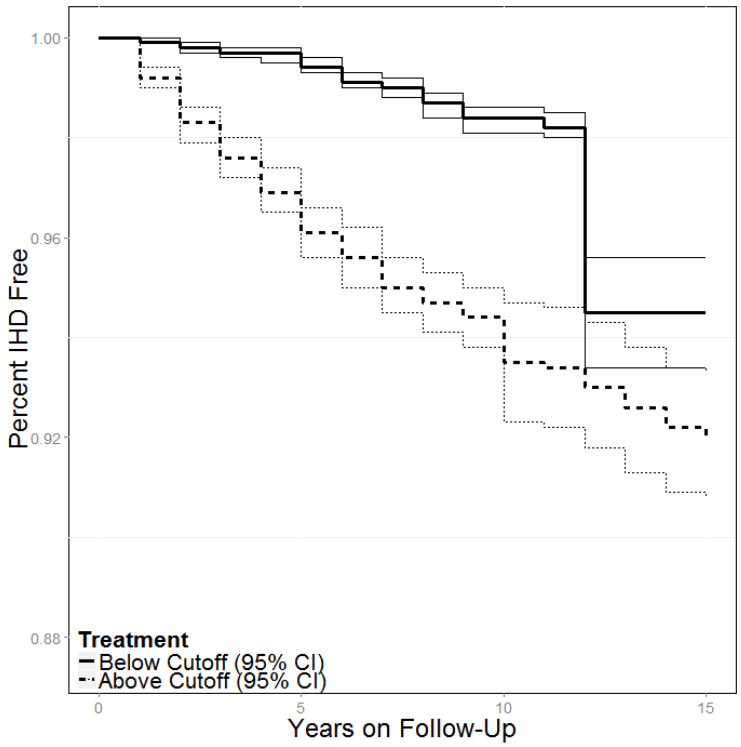

Figure 4.4: Estimated cumulative survival from IHD and 95% confidence intervals, adjusted for measured baseline and time-varying risk factors, among the fabricator worker population if continuously exposed vs unexposed at the 10th percentile cut-off of 0.06 $\frac{\text{mg}}{\text{m}^3}$

Chapter 5

Summary

In chapter 2, we argued for a broader understanding of the healthy worker survivor effect that incorporates the ideas of left truncation in the presence of heterogeneity in susceptibility as well as that of time-varying confounding affected by prior exposure. We have demonstrated that these phenomena result in a negative bias in effect estimates relative to those that would have been estimated in the full incident cohort. We have also demonstrated the viability of using well defined treatment regimens to analyze data from occupational cohorts where follow-up extends past employment termination.

In chapter 3, we explored the performance of a class of estimators of full data parameters of data structures in which some of the confounders are unmeasured, which we refer to as augmented inverse probability of censoring weighted targeted minimum-loss based estimators. Implementation of these estimators involve a targeting step for the fit of the conditional probability of missingness which can be approached in several ways. These estimators reduce bias and increase efficiency compared to the IPCW TMLE approach that they are based upon, with a minimal increase in computational complexity. Simulation demonstrated that the gains due to augmentation increase with the probability of missingness and that the double robustness property reduced bias compared to IPCW and MI approaches under misspecification of the estimator components Π_n and γ_n. The estimators were implemented for a longitudinal target parameter estimating the effect of occupational exposure to $PM_{2.5}$ on incidence of IHD in an aluminum smelter workforce.

In chapter 4, we demonstrated the application of TMLE in a longitudinal setting to account for time-varying confounding and generated doubly-robust, efficient, substitution estimators of our parameters of interest. These parameters were used to create marginal incidence curves that estimate the experience of the workforce if subjected to realistic interventions on the exposure assignment and censoring mechanisms. We believe that our analysis provides strong evidence of a causal connection between an accumulation of occupational exposure to $PM_{2.5}$ and the subsequent incidence of ischemic heart disease.

This dissertation considered topics in occupational epidemiology through the lens of causal inference. We used directed acyclic graphs and non-parametric structural equations to characterize our systems of interest and incorporated these structures into our definition and estimation of effect estimates. We recommend that occupational researchers follow the causal roadmap as set forth by van der Laan and Rose (2011), and, enumerate their assumptions about their system, suggest a target parameter of interest, evaluate its identifiability, and only then proceed with estimation.

Bibliography

Applebaum, K. M., Malloy, E. J., and Eisen, E. A. (2011). Left truncation, susceptibility, and bias in occupational cohort studies. *Epidemiology*, 22(4):599–606.

Arrighi, H. M. and Hertz-Picciotto, I. (1994). The evolving concept of the healthy worker survivor effect. *Epidemiology (Cambridge, Mass.)*, 5(2):189–96.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–73.

Bareinboim, E. and Pearl, J. (2012). Transportability of causal effects: Completeness results. In *AAAI*.

Bembom, O. and van der Laan, M. J. (2007). A practical illustration of the importance of realistic individualized treatment rules in causal inference. *Electronic journal of statistics*, 1:574–596.

Brookmeyer, R., Gail, M. H., and Polk, F. (1987). The prevalent cohort study and the acquired immunodeficiency syndrome. *American journal of epidemiology*, 126(1):14–24.

Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):571–584.

Chevrier, J., Picciotto, S., and Eisen, E. A. (2012a). 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. *Epidemiology (Cambridge, Mass.)*, 23(2):212–9.

Chevrier, J., Piciotto, S., and Eisen, E. A. (2012b). 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. *Epidemiology*, 23(2):212–219.

Choi, B. C. (1992). Definition, sources, magnitude, effect modifiers, and strategies of reduction of the healthy worker effect. *Journal of Occupational and Environmental Medicine*, 34(10):979–988.

Cole, S. R., Li, R., Anastos, K., Detels, R., Young, M., Chmiel, J. S., and Munoz, A. (2004). Accounting for leadtime in cohort studies: evaluating when to initiate hiv therapies. *Statistics in medicine*, 23(21):3351–3363.

Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D., and Poole, C. (2010). Illustrating bias due to conditioning on a collider. *International journal of epidemiology*, 39(2):417–20.

Cole, S. R., Richardson, D. B., Chu, H., and Naimi, A. I. (2013). Analysis of Occupational Asbestos Exposure and Lung Cancer Mortality Using the G Formula. *American journal of epidemiology*.

Costello, S., Brown, D. M., Noth, E. M., Cantley, L., Slade, M. D., Tessier-Sherman, B., Hammond, S. K., Eisen, E. A., and Cullen, M. R. (2014). Incident ischemic heart disease and recent occupational exposure to particulate matter in an aluminum cohort. *Journal of Exposure Science and Environmental Epidemiology*, 24(1):82–88.

Dumas, O., Le Moual, N., Siroux, V., Heederik, D., Garcia-Aymerich, J., Varraso, R., Kauffmann, F., and Basagaña, X. (2013a). Work related asthma. A causal analysis controlling the healthy worker effect. *Occupational and environmental medicine*, 70(9):603–10.

Dumas, O., Le Moual, N., Siroux, V., Heederik, D., Garcia-Aymerich, J., Varraso, R., Kauffmann, F., and Basagaña, X. (2013b). Work related asthma. A causal analysis controlling the healthy worker effect. *Occupational and environmental medicine*, 70(9):603–10.

Eisen, E. A., Holcroft, C. A., Greaves, I. A., Wegman, D. H., Woskie, S. R., and Monson, R. R. (1997). A strategy to reduce healthy worker effect in a cross-sectional study of asthma and metalworking fluids. *American journal of industrial medicine*, 31(6):671–677.

Eisen, E. A., Picciotto, S., and Robins, J. M. (2006). *Healthy Worker Effect*, pages 987–988. John Wiley & Sons, Ltd.

Fang, S. C., Cassidy, A., and Christiani, D. C. (2010). A systematic review of occupational exposure to particulate matter and cardiovascular disease. *International Journal of Environmental Research and Public Health*, 7(4):1773–806.

Fox, A. J. and Collier, P. (1976). Low mortality rates in industrial cohort studies due to selection for work and survival in the industry. *British journal of preventive & social medicine*, 30(4):225–230.

Gilbert, E. (1982). Some confounding factors in the study of mortality and occupational exposures. *American journal of epidemiology*, 116(1):177–188.

Gill, R. D., Wellner, J. A., and Prstgaard, J. (1989). Non- and semi-parametric maximum likelihood estimators and the von mises method (part 1) [with discussion and reply]. *Scandinavian Journal of Statistics*, 16(2):pp. 97–128.

Greenland, S. and Finkle, W. D. (1995). A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *Am. J. Epidemiol.*, 142(12):1255–1264.

Greenland, S., Pearl, J., and Robins, J. M. (1999a). Causal diagrams for epidemiologic research. *Epidemiology (Cambridge, Mass.)*, 10(1):37–48.

Greenland, S., Robins, J. M., and Pearl, J. (1999b). Confounding and Collapsibility in Causal Inference. *Statistical Science*, 14(1):29–46.

Handel, B. R. (2011). Adverse Selection and Switching Costs in Health Insurance Markets: When Nudging Hurts. *Working Papers, National Bureau for Economic Research*.

Hernán, M. A., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology (Cambridge, Mass.)*, 11(5):561–70.

Hernán, M. a., Cole, S. R., Margolick, J., Cohen, M., and Robins, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and drug safety*, 14(7):477–91.

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology (Cambridge, Mass.)*, 15(5):615–25.

Hernán, M. A. and VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology (Cambridge, Mass.)*, 22(3):368–77.

Howards, P. P., Hertz-Picciotto, I., and Poole, C. (2006). Conditions for Bias from Differential Left Truncation. *American Journal of Epidemiology*, 165(4):444–452.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4):523–539.

Kenward, M. G. and Carpenter, J. (2007). Multiple imputation: current perspectives. *Statistical methods in medical research*, 16(3):199–218.

Kenward, M. G. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, 13(3):236–247.

Klebanoff, M. A. and Cole, S. R. (2008). Use of multiple imputation in the epidemiologic literature. *American journal of epidemiology*, 168(4):355–7.

Kubo, J., Goldstein, B. A., Cantley, L. F., Tessier-Sherman, B., Galusha, D., Slade, M. D., Chu, I. M., and Cullen, M. R. (2013). Contribution of health status and prevalent chronic disease to individual risk for workplace injury in the manufacturing environment. *Occupational and environmental medicine*.

L Schafer, J. and W Graham, J. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2):147–177.

Le Cam, L. M., Le Cam, L. M., Le Cam, L. M., Mathematician, S., and Le Cam, L. M. (1986). *Asymptotic methods in statistical decision theory*. Springer-Verlag New York.

Lendle, S. D., Petersen, M. L., Schwab, J., and van der Laan, M. J. (2014). ltmle: An R Package Implementing Targeted Minimum Loss-based Estimation for Longitudinal Data. *Journal of Statistical Software (in review)*.

Li, L., Shen, C., Li, X., and Robins, J. M. (2011). On weighting approaches for missing data. *Statistical methods in medical research*, pages 0962280211403597–.

Little, R. and Hyonggin, A. (2003). Robust Likelihood-based Analysis of Multivariate Data with Missing Values.

Modrek, S. and Cullen, M. R. (2012). Job Demand and Early Retirement. *Working Papers, Center for Retirement Research at Boston College*.

Modrek, S. and Cullen, M. R. (2013). Health consequences of the Great Recession on the employed: Evidence from an industrial cohort in aluminum manufacturing. *Social Science & Medicine*, 92:105–113.

Monson, R. R. (1986). Observations on the healthy worker effect. *Journal of Occupational and Environmental Medicine*, 28(6):425–433.

Moore, K. and van der Laan, M. (2007). Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation.

Naimi, A. I., Cole, S. R., Hudgens, M. G., and Richardson, D. B. (2014). Estimating the effect of cumulative occupational asbestos exposure on time to lung cancer mortality: using structural nested failure-time models to account for healthy-worker survivor bias. *Epidemiology (Cambridge, Mass.)*, 25(2):246–54.

Naimi, A. I., Cole, S. R., Westreich, D. J., and Richardson, D. B. (2011). A comparison of methods to estimate the hazard ratio under conditions of time-varying confounding and nonpositivity. *Epidemiology (Cambridge, Mass.)*, 22(5):718–23.

Naimi, A. I., Richardson, D. B., and Cole, S. R. (2013). Causal inference in occupational epidemiology: accounting for the healthy worker effect by using structural nested models. *American journal of epidemiology*, 178(12):1681–6.

Nielsen, S. F. (2003). Proper and Improper Multiple Imputation. *International Statistical Review*, 71(3):593–607.

Noth, E. M., Dixon-Ernst, C., Liu, S., Cantley, L., Tessier-Sherman, B., Eisen, E. A., Cullen, M. R., and Hammond, S. K. (2013). Development of a job-exposure matrix for exposure to total and fine particulate matter in the aluminum industry. *Journal of exposure science & environmental epidemiology*.

Ogle, W. (1885). *Annual Report of the Registrar-General for England and Wales*, volume 45. HM Stationery Office.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Pearl, J. and Bareinboim, E. (2011). Transportability of Causal and Statistical Relations: A Formal Approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 540–547. IEEE.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379.

Petersen, M., Porter, K., Gruber, S., Wang, Y., and van der Laan, M. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31.

Petersen, M. L. (2011). Compound treatments, transportability, and the structural causal model: the power and simplicity of causal graphs. *Epidemiology (Cambridge, Mass.)*, 22(3):378–81.

Picciotto, S., Brown, D. M., Chevrier, J., and Eisen, E. A. (2013). Healthy worker survivor bias: implications of truncating follow-up at employment termination. *Occupational and environmental medicine*, 70(10):736–742.

Picciotto, S., Chevrier, J., Balmes, J., and Eisena, E. A. (2014). Hypothetical interventions to limit metalworking fluid exposures and their effects on copd mortality. *Epidemiology*, 20:00–00.

Pope III, C. A. (2002). Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. *JAMA: The Journal of the American Medical Association*, 287(9):1132–1141.

R Core Team (2013a). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2013b). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Richardson, D., Wing, S., Steenland, K., and McKelvey, W. (2004). Time-related aspects of the healthy worker survivor effect. *Annals of epidemiology*, 14(9):633–639.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.

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

Robins, J. (2000a). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, pages 6–10.

Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese*, 121(1):151–179.

Robins, J. M. (2000b). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.)*, 11(5):550–60.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, 90(429):122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):pp. 846–866.

Ronneberg, A. (1995). Mortality and cancer morbidity in workers from an aluminium smelter with prebaked carbon anodes–Part III: Mortality from circulatory and respiratory diseases. *Occupational and Environmental Medicine*, 52(4):255–261.

Rose, S. and van der Laan, M. J. (2012). A targeted maximum likelihood estimator for two-stage designs. *The international journal of biostatistics*, 7(1):17.

Rotnitzky, A. and Robins, J. M. (1995). Semi-parametric estimation of models for means and covariances in the presence of missing data. *Scandinavian Journal of Statistics*, 22(3):pp. 323–333.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley.

Rubin, D. B. (1976a). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1976b). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91(434):473.

S Su, Y., Gelman, A., Hill, J., and Yajima, M. (2011). Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. *Journal Of Statistical Software*, 45(2):1–31.

SAS Institute Inc. (2011). *SAS/STAT 9.3 User's Guide*. SAS Institute Inc, Cary, NC.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, 1 edition.

Schnitzer, M. E., Moodie, E. E., and Platt, R. W. (2013). Targeted maximum likelihood estimation for marginal time-dependent treatment effects under density misspecification. *Biostatistics*, 14(1):1–14.

Schwab, J., Lendle, S., Petersen, M., and van der Laan, M. (2013). *ltmle: Longitudinal Targeted Maximum Likelihood Estimation*. R package version 0.9.1.

Seaman, S. R. and White, I. R. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*.

Stayner, L., Steenland, K., Dosemeci, M., and Hertz-Picciotto, I. (2003). Attenuation of exposure-response curves in occupational cohort studies at high exposure levels. *Scandinavian Journal of Work, Environment and Health*, 29(4):pp. 317–324.

Steenland, K. (2013). Marginal structural models to control for time-varying confounding in occupational and environmental epidemiology. *Occupational and environmental medicine*, 70(9):601–602.

Steenland, K., Deddens, J., Salvan, A., and Stayner, L. (1996). Negative bias in exposure-response trends in occupational studies: modeling the healthy worker survivor effect. *American journal of epidemiology*, 143(2):202–210.

Stitelman Ori, M., van der Laan Mark, J., et al. (2012). A general implementation of tmle for longitudinal data applied to causal inference in survival analysis. *The International Journal of Biostatistics*, 8(1):1–39.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer.

van der Laan, M. and Gruber, S. (2011). Targeted Minimum Loss Based Estimation of an Intervention Specific Mean Outcome.

van der Laan, M., Polley, E., and Hubbard, A. (2007). Super Learner.

van der Laan, M. J. and Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international journal of biostatistics*, 8(1).

van der Laan, M. J. and Petersen, M. L. (2007). Causal Effect Models for Realistic Individualized Treatment and Intention to Treat Rules. *The international journal of biostatistics*, 3(1):Article3.

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer.

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning*. Springer.

van der Laan, M. J. and Rubin, D. (2006a). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, 2(1).

van der Laan, M. J. and Rubin, D. (2006b). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, 2(1).

Wang, M.-C., Brookmeyer, R., and Jewell, N. P. (1993). Statistical models for prevalent cohort data. *Biometrics*, pages 1–11.

Westreich, D., Cole, S. R., Schisterman, E. F., and Platt, R. W. (2012). A simulation study of finite-sample properties of marginal structural Cox proportional hazards models. *Statistics in medicine*, 31(19):2098–109.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399.

Williamson, E. J., Forbes, A., and Wolfe, R. (2012). Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder. *Statistics in medicine*, 31(30):4382–400.

5.1 Introduction

The healthy worker survivor effect has been a known problem for occupational researchers for many years. As the fields of epidemiology, biostatistics, and causal inference have developed in this time, so too have the field's understanding of the HWSE and the proposed methods for properly adjusting for it. In this appendix we describe select papers that represent key contributions to this process. This historical perspective highlights the great strides made in both the philosophy and methodology of occupational health over this time.

5.2 Literature Review

Fox and Collier (1976) is the first modern reference that most researchers cite for the healthy worker effect, although the earliest recorded mention of it was by William Ogle in the late 1800's (Ogle, 1885). Fox and Collier identified three separate aspects of the healthy worker effect. The first aspect was the selection effect, in which people selected into a workforce are healthy and robust enough to work and therefore have lower mortality than the general population. The second aspect was the survival effect, in which workers who continue to work are healthier than a comparable group that terminates employment. The third aspect was the length of follow up effect, in that workers with shorter work histories tend to to be less stable an therefore at a higher risk of death. The selection effect is now generally referred to as the healthy hire effect, while the second and third aspects are portions of what is generally referred to as the healthy worker survivor effect (HWSE).

Fox and Collier demonstrated the healthy worker effect using standardized mortality ratios (SMRs), which the mortality experience of a cohort is indirectly compared with the experience of the general population, adjusting for age, sex, race, and calendar year. They chose different cohorts of the data to illustrate each of the aspects of the healthy worker effect, which they identified with SMRs under 100. For the healthy hire effect, they looked at the effect of the length of time since the start of work, demonstrating that the reduction in SMR is almost eliminated after 15 years since start of hire. For the healthy worker survivor effect, they demonstrated reduced SMRs among active workers, and elevated SMRs among their inactive counterparts, 15 years following the initial hire date.

Following Fox & Collier's illustrative paper, many occupational analyses were done using internal comparison groups, controls who were employed in similar industrial jobs. However, this correction is not sufficient to estimate the true effect of exposure, because some portions of the working population may be more affected by the HWE than others. Specifically, workers that maintain health as well as employment may be likely to take jobs with the highest exposure levels, due to either their robustness or seniority (highly exposed jobs are often demanding and dangerous and command a higher salary).

Gilbert (1982) argued that SMRs were not an adequate effect measure for occupational exposures, and proposed an alternative methodology based on the Mantel Haenszel procedure. She highlighted several inefficiencies with SMRs: they do not adequately adjust for confounding by age, do not accurately represent the underlying risk ratios and require exponentially more data as the number of confounders increase. Gilbert's most lasting contribution was twofold. The first was her recommendation that short-term workers be removed from internal comparisons, as they are possibly not exchangable with the rest of the population. Second, she recommended the introduction of lagged exposure variables, which can reduce the HWSE in diseases with long latencies.

Nonetheless, Monson (1986) used SMRs below 100 as the primary evidence for the existence of the HWE. In an exhaustive analysis, he selected a number of potential confounders through which to examine the HWSE. One of his most important observations was that the HWE, as demonstrated through SMR reduction, seems to have a dynamic phase earlier in follow-up, during which the the healthy hire effect disapates as the employed workforce begins to more closely resemble the general workforce. Subsequent to this, he observed a plateau phase later on, in which the underlying differences between an employed and general population remain observable and unchanging. He observed a longer dynamic phase of 30 years than the 15 observed by Fox and Collier, probably due to his restriction to employed workers (no follow-up after employment termination). Monson's ultimate view is that it is primarily due to two overriding factors. The first is the selection bias from hiring especially healthy people. The second is the underlying demographic differences between the working and general populations.

In 1987 Robins introduced a new approach to thinking about the healthy worker survivor effect (Robins, 1987). He was the first to frame the HWSE as a bias that causes the true effect of exposure on outcomes to be underestimated. This notion of a causal parameter as a target of study was relatively novel at this point, and necessitated a further level of formality in describing the relationships between exposure, employment status, and morbidity. As a tool to accomplish this, Robins introduced measured graphs, in which each time point was represented by a node, and nodes were connected by paths. The paths represented different combinations of exposure level and employment status, and split and inter and extra-nodal points. The difference between the two path splitting points is the notion of exchangability. Extra-nodal splits represent populations splitting on exposure levels must be assumed to be exchangable in order for a causal effect to be identifiable. Intra-nodal splits, by contrast, represent subjects leaving work and the populations split between them do not need to satisfy this same assumptions in order for the effect to be identifiable. This is because, in a hypothetical randomized trial on exposure status in a workplace, people who leave work would be no longer participating in the trial. They must be accounted for, but can be assumed to have a different mix of confounders from the workers who remain. Robins used these graphs to define the G-causal parameter, one of the first examples of an unbiased estimator of the effect of exposure on mortality that appropriately adjusts for the

intermediate variable of employment history.

Robins approach represented a new methodology for adjusting for confounding that could take account of longitudinal exposures affecting future employment history which in turn affects future exposure. This time-varying covariate, which is also possibly intermediate to the outcome of interest, is not properly controlled for using standard methodologies of confounding adjustments. This confounder is certainly one of the primary drivers of the HWE and a part of the reason why it has been so difficult to correct for. Robins proposed two methodologies for dealing with the HWE in occupational cohort analysis: the G computation algorithm, the G null test and G-estimation. The G-computation algorithm is method which uses fits of the successive regressions that generate the observed data to estimate the experience of a cohort subject to a specific history of exposure and work status. The G-null test is a proposed methodology for testing the the hypothesis that all causal parameters are identically equal to 0.

G-estimation is a method for estimating the parameters of structural nested failure time models, and it has the useful property of being able to utilize data from worker follow-up after employment termination. It proceeds by first generating a model for $g(t)$, the probability of being exposed at time point t, given a worker's history. This probability is obviously 0 for all workers after they are terminated. G-estimation proceeds by then using a range of values for the parameters of the failure time model to generate estimates of T_0, the time of each worker's failure if never exposed. Each of these vectors of times are then checked for independence from the probability of exposure, and the estimate is the parameter value that adds no information to the g model. This approach works by leveraging the no unmeasured confounders assumption that failure time is independent from exposure assignment at all time points, given the observed data. While Robins ideas represented a sea change in their approach to the HWE, they were not recognized by the occupational health community for the next fifteen to twenty years (Arrighi and Hertz-Picciotto, 1994) and have only recently been applied to real data (Chevrier et al., 2012a, Naimi et al., 2014, Picciotto et al., 2014).

These ideas did not catch on within the world of occupational epidemiology immediately. Indeed, a review article from 1992 Choi (1992), in which Choi interviewed nine occupational epidemiologists to gain a consensus on the definition, sources, and strategies for the reduction of the Healthy Worker Effect, there is no mention of Robin's work. The only point of agreement between all interviewees was that the HWE is 'an observed decrease in mortality among workers', in relation to the general population. Four generalized sources of the HWE are identified, though: the true effect of work on health, selection bias, confounding bias, and information bias. The confounding and selection biases both primarily deal with differential health statuses between workers and non-workers. Information bias arises because individual workers must be positively identified by a death record, while the general population rates are summarized, and therefore missing matches may underestimate the true death rate of workers. A variety of techniques were propose to adjust for the effect, representing the divergent views of researchers at the time.

One new strategy for adjusting for the HWE was proposed by Eisen et al (Eisen et al., 1997). The target aspect of this analysis was the transfer bias, in which people at higher risk for the disease of interest reduce their exposure not by leaving work, but by transferring to lower exposed positions within the workplace. This method uses a cross-sectional sample of a working population that has full information on job transfer, exposure and disease status. The population is then analyzed ysing a longitudinal model, as if every person was followed from their date of hire, and case exposure is measured as it was prior to disease onset. For a rapid onset disease such as asthma, this method can reduce negative bias due to asthmatics moving out of jobs with breathing exposures. This method was demonstrated to increase the exposure-response effect estimate, but does not account for additional bias caused by transfer out of the company altogether.

Steenland and Stayner (Steenland et al., 1996) were among the first to investigate the healthy worker effect using simulated data sets. They used cumulative measure of the exposure of interest and simulated an increased mortality rate among workers following employment termination. As predicted by theory, this resulted in significant protective effects of cumulative exposure when modeling the dose-response for exposure and mortality. This was observed when the true effect of exposure was both positive and null. They were able to somewhat reduce the bias by including the active status of an employee as an interaction term with exposure in the model. They discuss how the inclusion of active status as does not fully account for its role as a mediator of the exposure-disease response, but postulate that this effect is small compared to work statuses effect as a confounder.

Prior to 2004, one aspect of the healthy worker effect that had not received very much attention by researchers was the change in mortality rates following termination. Workers who contract a deadly disease often have the opportunity to quit work prior to the disease running its course. Richardson et al. (2004) demonstrated an increase in mortality subsequent to employment termination using several empirical data sources. They augment these observations with simulation analysis to quantify the effect on the exposure-response estimates. The simulation demonstrated that this effect alone, in the absence of any effect of exposure on work status or other time-varying covariates, can cause a negative bias. However, this aspect can be eliminated by adjusting for time since termination in a proportional hazards model.

Another important feature of the HWE that is more recently being appreciated is the role that disease susceptibility can play in its formation. A recent paper by Applebaum et al (Applebaum et al., 2011) used simulation to demonstrate that bias is induced in effect estimates when the analysis cohort is left truncated. Left truncated cohorts consist of prevalent hires, workers who have survived for some amount of time between hire and follow-up start. If an occupational cohort consists of workers with a range of susceptibility to the exposure of interest, a prevalent cohort will contain fewer of the more susceptible workers. This results in a downwards bias in effect estimates compared to the true effect in an incident cohort, which is often the true target parameter. The authors suggest that using time since hire as

a time metric prevent this problem, and suggest the use of flexible modeling approaches to examine the changes in the true hazard ratio over time.

5.3 Discussion

Over the past 35 years, the understanding of the effect of the HWE on studies in occupational health has been greatly enriched by the work of the researchers whose work is listed above, as well as many others. What seems on the surface to be a relatively simple combination of two selection factors has shown itself to contain many aspects that are not immediately apparent. In order to arrive at a cohesive theory of the HWE, several of these aspects deserve special attention.

In Fox and Collier's original paper, they noted that the HWE seemed to wear off after 15 years, and this statement has been often repeated in the literature. While true in their observation, it relied on the use of an SMR which compares to the general population as an outcome measure and on cohort members continuing on follow-up past employment termination. As time-since-follow-up increases, a larger percentage of the initial study population are not longer active, and eventually represent a similar mix of employment statuses as the general population. The estimate of 15 years is dependent on the rate of leaving work within the cohort, which we would not expect to be constant across time and different industries. Monson's analyses, which stratified on employment status, confirms this view, as he observed reduced SMRs as far as 30 years out from the hire date. From a causal perspective, work and health form a continuous loop, with work providing the financial, medial, and psychological support necessary to maintain personal health, while this health allows a person to continue working. What is clear is that as time moves forward from the start of hire, the healthy hire effect will continue to reduce its effect, but the survivor effect will continue.

While consideration of the HWE is a necessary step for any study involving an occupational exposure, the different understandings and approaches to controlling for it suggested by the authors we discuss illustrate an important point; its effects are highly specific to the exposure, outcome, industry, and study design. For instance, the lagging of an exposure may make sense for a disease with a long latency period, such as cancer, but is not a reasonable approach for heart disease. Additionally, concerns about controlling for time-since termination are only relevant in cases in which follow-up extends past a termination event. If follow-up ends at termination, than the possibility of differential censoring by health status must be addressed. Adjustment for censoring or exposure assignment necessitates a detailed understanding of the processes for hiring and job placement strategies, an appraisal of the immediate physical effects of exposure, and the other opportunities available to the workforce.

Many variables traditionally available in an occupational cohort study have been used to attempt to adjust for the HWE, including time-since-hire, active status, time-since-follow-

up-start, and time-since-termination. Controlling for each individually or all together can help to elucidate aspects of the effect and reduce bias as compared to an uncontrolled study. However, standard methods of covariate control, such as stratification, can not properly deal with time-varying confounding on the causal pathway. This is why we focused our analyses in this dissertation on methods that can handle such variables.