A Causally Formulated Hazard Ratio Estimation through Backdoor Adjustment on Structural Causal Model

Riddhiman Adib   RIDDHIMAN.ADIB@MARQUETTE.EDU
Department of Computer Science
Marquette University
Milwaukee, Wisconsin, USA

Paul Griffin   PAULGRIFFIN@PURDUE.EDU
Regenstrief Center for Healthcare Engineering
Purdue University
West Lafayette, Indiana, USA

Sheikh Iqbal Ahamed   SHEIKH.AHAMED@MARQUETTE.EDU
Department of Computer Science
Marquette University
Milwaukee, Wisconsin, USA

Mohammad Adibuzzaman   MADIBUZZ@PURDUE.EDU
Regenstrief Center for Healthcare Engineering
Purdue University
West Lafayette, Indiana, USA

Abstract

Identifying causal relationships for a treatment intervention is a fundamental problem in health sciences. Randomized controlled trials (RCTs) are considered the gold standard for identifying causal relationships. However, recent advancements in the theory of causal inference based on the foundations of structural causal models (SCMs) have allowed the identification of causal relationships from observational data, under certain assumptions. Survival analysis provides standard measures, such as the hazard ratio, to quantify the effects of an intervention. While hazard ratios are widely used in clinical and epidemiological studies for RCTs, a principled approach does not exist to compute hazard ratios for observational studies with SCMs. In this work, we review existing approaches to compute hazard ratios as well as their causal interpretation, if it exists. We also propose a novel approach to compute hazard ratios from observational studies using backdoor adjustment through SCMs and do-calculus. Finally, we evaluate the approach using experimental data for Ewing's sarcoma.

1. Introduction

Experimental studies such as randomized controlled trials (RCT) are considered the gold-standard in hypothesis testing. For safety and efficacy reasons and regulatory purposes, most new drugs or treatments are studied through RCTs (Greene and Podolsky, 2012). RCTs provide the best mechanism to identify the causal effect of treatments or interventions, by adjusting for observed and unobserved confounders under the rubric of a potential outcome framework (Fisher et al., 1960). Despite clear advantages of RCTs in drug-trials, in

practice, they are expensive, time-consuming, and not feasible in many cases due to ethical reasons. Other issues with RCTs include low recruitment rate, loss to follow-up, insufficient sample size, and being prone to selection bias (Nichol et al., 2010; Frieden, 2017). While RCTs remain the best way to establish causation, large amounts of data captured with new technologies during routine healthcare (e.g., electronic health records (EHR) or wearable devices), colloquially termed *big health data*, has the potential to discover causal effects from observational studies to complement RCTs. With proper methodological considerations, observational studies can provide a way to *emulate* RCTs and go beyond statistical correlation (Hernán et al., 2008; Hernán and Robins, 2017).

In the 1970s, the potential outcome framework was extended to observational studies to identify causal relationships from observational data through the Rubin Causal Model (RCM) (Rubin, 1974; Rosenbaum and Rubin, 1983; Holland, 1986). Recent advances in structural causal model (SCM) provides the methodological framework under the potential outcome framework for graphically formalizing the identification of causal effects from observational and experimental data (Pearl, 2009; Pearl et al., 2016). SCMs can be used to *emulate* RCTs from observational data in many cases if the graphical model is identifiable (Bareinboim and Pearl, 2016), which signifies the capability of estimating the interventional distribution ($P(y|do(x))$) from the available data with the assumptions incorporated in the model.

Experimental studies (including RCTs) frequently explore and report survival analysis measures. Survival analysis is the branch of statistics that analyzes the expected duration of time-to-event with outcome statistics such as hazard ratio, odds ratio, and risk ratio. Survival analysis has been well-studied under the potential outcome framework with experimental studies and with RCM for observational studies (Cole and Hernán, 2004; Hernán, 2010). Recent research has also studied survival analysis with RCM for observational studies considering the data generating mechanism or the study designs to estimate outcome statistics such as hazard ratio, odds ratio, risk ratio, and risk difference (Didelez and Sheehan, 2007; Hernán, 2004).

Commonly reported outcomes from survival analysis in experimental clinical studies include the survival curve and hazard ratio (HR). The survival curve graphically reports the hazard in a population and represents the fraction of the population that survived in the treatment and the control group over time. HR describes the comparative hazard between the treatment and the control group. Hazard function, or simply *hazard* signifies the rate of events-of-interest (e.g., a death) at time t, conditional on survival until time t and beyond (Spruance et al., 2004). For example, we present a survival curve (Figure 1) as reported in (Girard et al., 2018), where probability of overall survival of patients in drug groups (starting at 100%) is presented with time passed, and the probability declines with time.

Even though HR is widely used in practice as a standard tool for comparative evaluation of the outcome between treatment and control groups, it depends on the length of the study and, by definition, has an inherent selection bias (since only the survived population at time t are *selected* at time $t+1$) (Hernán, 2010). In addition, both the survival curve and HR do not consider the study design, that is RCT versus observational study, in their formalization. Consequently, it is difficult to interpret the results of an intervention from only the reported hazard ratio (Hernán, 2010) and compare different studies with varying study designs and time lengths. The researcher has to consider the design of the study, length of the study
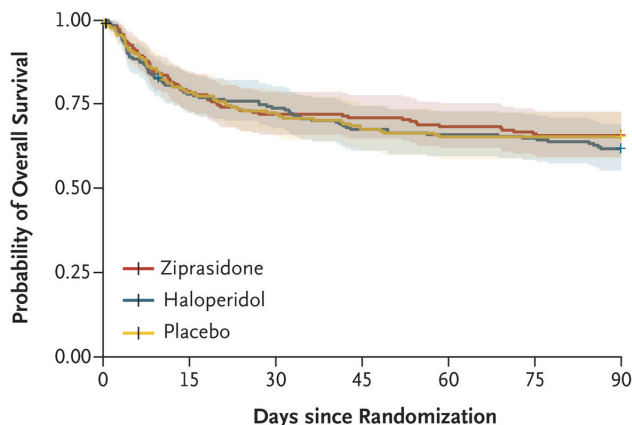
Figure 1: An example survival curve, collected from (Girard et al., 2018)

as well has the hazard ratio to understand the effectiveness of the treatment. Structural causal models (SCMs) provide a framework to explicitly define the design of the study, the assumptions for the study, as well as the length of the study. However, to the best of our knowledge, a framework to compute the hazard ratio with SCMs does not exist.

Previous approaches for adjusted survival curves under the rubric of RCM used inverse probability weighting (IPW) to adjust for confounders in the estimand (Cole and Hernán, 2004). However, this approach has a strong assumption, namely *ignorability* (Rubin, 1978; Angrist et al., 1996). The ignorability assumption states that there are no unobserved confounders in the model, and the variables considered for IPW satisfy the backdoor criterion. Although an approach with instrumental variable can be used when the treatment assignment is non-ignorable (Angrist et al., 1996), in practice, this is rather a strong assumption and a variable can be a mediator, a collider, an M-bias, or a confounder (Lederer et al., 2019). In this paper, we formulate the estimation of the hazard ratio from observational studies under the rubric of SCMs that does not depend on the ignorability assumption. We provide a principled approach to define observational studies using SCMs, redefine with time-specific survival as outcomes (instead of survival time as the only outcome), and therefore mathematically transform observational studies to the corresponding experimental studies by adjusting for confounders with the backdoor criterion and then, sample from the experimental studies to estimate hazard ratios. We provide the mathematical formalization of the approach with a simple causal graph and with detailed mathematical derivation, and validate the results with a simulated data set and a benchmark data set on Ewing's sarcoma.

1.1. Clinical Relevance

Most clinical research reports HR with survival analysis. However, the reported HR and its process of calculation do not take into account the study design (e.g., RCT vs. observational study) and corresponding assumptions (e.g., ignorability). This makes it harder to compare the results of different studies with different study designs, sample populations, study lengths and assumptions. Our proposed method with SCMs estimates HR by explic-

itly describing the study designs and assumptions for a better clinical understanding of the effect of the treatment of interest.

1.2. Technical Significance

We propose a novel approach to estimate the HR from observational studies with SCM, taking the causal relationship between treatment and outcome into account. In HR calculation for survival analysis of observational studies, our review of the literature identifies a lack of causal interpretation. Our proposed approach first develops a time-invariant causal model and estimates the survival time after adjusting for the confounders in the SCM using backdoor adjustment and do-calculus. The development of an SCM enables us to identify the confounding variables, unlike with the ignorability assumption where we adjust for every variable available (except treatment and outcome), as well as properly adjust using the minimal set, thus reducing computational requirements. The computed survival times are considered "as-if" they were sampled from an RCT. The newly adjusted survival times are capable of expressing the true causal effect of treatment on the outcome through the survival curve and HR. We validate the proposed method in both simulated experiments and with observational data.

1.3. Generalizable Insights

We propose a novel method of estimating the HR for observational studies under the rubric of SCMs. The method can be used for any observational studies with survival data, after defining the SCM. Our method of estimating the HR through SCMs clearly defines the study-design and assumptions in the model. All the source code for this study is shared with the research community through a GIT repository. A Python-based library has been released that takes the data, the graph, and length of the study as input and provides the adjusted survival curve with backdoor adjustment and the hazard ratio as the output. Our approach is limited in the cases when i) the SCM is not defined and ii) the SCM is not identifiable through the adjustment formula or backdoor adjustment (i.e., there is no backdoor set).

2. Related Work

Survival analysis (Kleinbaum and Klein, 2010) is a methodological approach for modeling and comparing the time-to-event between two populations. The event is called a hazard, which can be death, an adverse clinical event, or a mechanical failure for physical systems. It compares the condition of survival in the treatment versus control group, and reports outcomes with statistical measures such as the HR. Frequently reported approaches in survival analysis include Kaplan Meier survival curve, Cox proportional hazards model, life tables, and survival trees,. We review a non-parametric approach of the Kaplan Meier survival curve and the semi-parametric approach of the Cox proportional hazards model.

The Kaplan Meier survival curve (Kaplan and Meier, 1958) is a non-parametric statistic representing the survival function and HR in the treatment and the control group. It provides a visual comparison between survival functions in different treatment or control groups; it does not differentiate between RCTs or observational studies. Data from both

of the approaches can be plotted as the Kaplan Meier curve. It is up to the individual researcher to interpret and explain the Kaplan Meier curve based on the study design. Cox PH model, on the other hand, is computationally complex. However, it is a commonly used approach for survival analysis, and is widely used to compute the HR in epidemiological studies. The key aspect of it is the underlying proportional hazards assumption (Cox, 1972), stating that the HRs of the treatment and control group are proportional and is a function of the covariates. It is a semi-parametric model since no assumption is made about the baseline hazard function (i.e., hazard function with no covariates). In general, it is effective in estimating both regression coefficients (β_i) and the HR (Kleinbaum and Klein, 2010). Futher, it is unbiased *(when estimated considering all possible covariates)*.

We review existing approaches to compute the HR for observational and experimental studies. Previous work on survival analysis for observational data with RCM under potential outcome used IPW to adjust for confounders (Cole and Hernán, 2004). However, RCM requires the ignorability assumption that all variables considered for adjustment with IPW satisfy the backdoor criterion. In reality, a variable can also be a mediator, and in those cases adjusting for the mediators will result in inaccurate analysis. It has also been shown that the HR estimation approach has an inherent selection bias (Hernán et al., 2004; Hernán, 2010) as only the patients who survived at time t were sampled at time $t + 1$ to be considered for the estimation. SCMs provide the mathematical machinery to identify the backdoor variables given a causal graph. We used the same Ewing's sarcoma data set as studied in (Cole and Hernán, 2004) with the same assumptions (i.e., all the covariates satisfy the backdoor criterion) to arrive at the same result as a validation strategy for our approach.

For survival analysis, it was shown that in some cases the Kaplan Meier curve may show no difference between the treatment and control groups when in reality there is a statistically significant difference in the HR, if it is adjusted properly (Makuch, 1982). The rationale behind this phenomenon is that a non-parametric approach is used to plot the survival curve, whereas a semi-parametric method is used to calculate the HR. The authors (Makuch, 1982) presented an approach to construct a plot of the survival curve consistent with the HR calculated. In this work, the adjusted survival curve for a specific treatment group was estimated by calculating a mixture of the estimated survival functions for separate strata, and weighted based on the distribution of the covariate in the sample dataset. However, the approach does not consider the design of the study in the survival analysis.

To extend the existing definition of the Cox PH Model, the Marginal Structural Cox PH Model has been introduced and used to find the effect of Zidovudine on the survival of HIV-positive men (Hernán et al., 2000). Statistical analysis in the presence of time-dependent confounders is commonly done through a standard Cox PH model. However, Robin (Robins, 1997) has previously shown that this approach cannot adjust for all biases. Similar to previous work under the RCM, the authors used the conditional ignorability assumption. This is a much stronger assumption compared to using the SCM to identify confounding variables opening the backdoor. Several other researchers (Satten and Datta, 2001; Rotnitzky and Robins, 2014) have used the IPW approach, although without using SCMs.

The existing literature to compute the HR does not consider the study design and might lead to misinterpretation if the data were not sampled correctly or adjusted for the

right confounding variables. While previous research alludes to this problem, they do not provide the mathematical machinery for survival analysis. Although the traditional Cox PH Model can minimize the effects of biases, it is not the same as "adjustment" of confounding variables. The bias is reduced by fitting the Cox PH regression model until convergence (Cox, 1972), it does not consider the study design or the data generating mechanism. The model fitting approach does not generate a causally meaningful interpretation despite reduction in biases. Our goal is to formulate an approach that estimates the HR through a causal formulation considering the data generating mechanism with SCM, that portrays the direct causal effect of treatment on outcome, in terms of the HR metric. The assumption of variables opening the backdoor path in the SCM as confounders and adjustment on the dataset based on that enables a more causally interpretable estimation of the HR.

3. Background

3.1. Hazard Ratio

To define the HR, we use the hazard function (Spruance et al., 2004) in the Cox proportional hazard model:

$$h(t, \mathbf{X}) = h_0(t) \exp(\sum_{i=1}^{p} \beta_i X_i) \tag{1}$$

Based on this, the Hazard Ratio (HR) is defined (Kleinbaum and Klein, 2010) as:

$$HR = \frac{h(t, \mathbf{X}_{x=1})}{h(t, \mathbf{X}_{x=0})} \tag{2}$$

Here, $h(t, \mathbf{X})$ represents the hazard function at time t and the vector with the covariates of the model $\mathbf{X}$. $\mathbf{X}$ can also be written as $[w_0, w_1, ..., w_m, z_0, z_1, ..., z_n, x]$, where x is the treatment, z_i are the confounders, and w_i are the other associated covariates. $\mathbf{X}_{x=1}$ represents the value of the covariate vector $\mathbf{X}$ with value of the treatment set as 1 ($x = 1$), making $\mathbf{X}_{x=1} = [w_0, w_1, ..., w_m, z_0, z_1, ..., z_n, 1]$. β represents the *maximum likelihood estimates (MLE)* for each covariate. In other words, β is the corresponding coefficient for each covariate that fits the data into a converging model for the Cox regression.

As expressed in Equation 1, the proportional hazard assumption defines the hazard function $h(t, \mathbf{X})$ to be composed of the baseline hazard function $h_0(t)$ (i.e., hazard when all covariates are set to 0), multiplied with the exponential of the sum of β multiplied by the corresponding covariate.

Since we have defined the HR and hazard function, we can simplify the equation of the HR to:

$$\begin{aligned} HR &= \frac{h(t, \mathbf{X}_{x=1})}{h(t, \mathbf{X}_{x=0})} \\ &= \frac{h_0(t) \exp(\beta_x 1 + \beta_z Z + ...)}{h_0(t) \exp(\beta_x 0 + \beta_z Z + ...)} \\ &= \exp(\beta_x) \end{aligned} \tag{3}$$

6

In other words, the HR is equivalent to the exponential of the regression coefficient β. However, computing β is non-trivial since, in practice, one does not know the baseline hazard function $(h_0(t))$. We can only estimate the HR using the maximum likelihood function, and iterating until the model converges to a pre-defined error bound (Kleinbaum and Klein, 2010).

Although the HR is an important outcome, it has limitations in explaining causal relationships. No causal mechanism is understood from the HR. This is because the HR is calculated from the convergence of regression models and, confounding and other such bias is handled by simply including the covariates to the model. It is then up to the individual researcher to make sure that the right data are used to measure the HR and interpret accordingly. For example, an HR calculated from an RCT provides the casually linked hazard for the intervention, whereas the same HR calculated from an observational study simply provides a correlated hazard. This existing approach simplifies the calculation and reduces the burden on the researcher. However, we frequently find differences between the survival curve and the HR. This difference, or bias, arises because of the inherent definitions of the survival curve and Cox PH model.

3.2. Structural Causal Models

Structural causal models (SCMs), developed on the foundations of probabilistic graphical models, draw inferences that explain the causal relationship between variables. With an SCM, a causal model is defined first and is expressed with a graphical representation. Definition 1 gives the formal description of an SCM: (Bareinboim and Pearl, 2016; Pearl, 2009).

Definition 1 (Structural Causal Model) *A structural causal model M is a 4-tuple $\langle U, V, f, P(u) \rangle$ where:*

1. *U is a set of background (exogenous) variables that are determined by factors outside of the model,*

2. *V is a set $\{V_1, V_2, ..., V_n\}$ of observable (endogenous) variables that are determined by variables in the model (i.e., determined by variables in $U \cup V$),*

3. *F is a set of functions $\{f_1, f_2, ..., f_n\}$ such that each f_i is a mapping from the respective domains of $U_i \cup PA_i$ to V_i, where $U_i \subseteq U$ and $PA_i \subseteq V$ V_i and the entire set F forms a mapping from U to V. In other words, each f_i in $v_i \leftarrow f_i(pa_i, u_i), i = 1, ..., n$, assigns a value to V_i that depends on the values of the select set of variables $(U_i \cup PA_i)$, and*

4. *$P(u)$ is a probability distribution over the exogenous variables.*

An SCM is often expressed by a causal graph G. Each node V in G represents an observed or unobserved variable, and each directed edge E represents the causal relationships between them. To find the causal effect of variable X on variable Y, do-calculus is introduced (Pearl et al., 2016). Do-calculus is used to map the observational reality to the corresponding experimental reality with the identifiability equation by adjusting for different kinds of biases (e.g., confounding bias), if it exists. The backdoor criterion provides a

powerful tool to identify the variables that need to be adjusted for this transformation (in other words, adjust for confounding bias) and is defined in definitions 2 and 3.

Definition 2 (Backdoor Criterion) *Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.*

Definition 3 (Backdoor Adjustment) *If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by the formula:* $P(y|do(x)) = \sum_z P(y|x, z)P(z)$

3.3. Problem Definition

Our research problem is to develop a method to compute the HR for observational studies by leveraging the SCM by explicitly declaring our assumptions and adjusting for the right confounders. The goal is to acknowledge the defined roles of variables in the SCM, and use a minimum set of confounders to adjust for backdoor, thus building a computationally-efficient and more accurate model for objective estimation and comparison. The algorithm will take three sets/inputs, (1) observational dataset D consisting of treatment, outcome in survival-time and other covariates, (2) SCM supporting the causal mechanism and dataset, G, and, (3) length-of-trial T. At the completion of the algorithm, the output will be: (1) adjusted survival curve S *(non-parametric estimation)*, and (2) hazard ratio of treatment, HR *(semi-parametric estimation)* (Figure 2). The assumption in our approach is that the observational data are available, and the SCM is fully specified.

4. Methods

In this section, we formalize our approach to mathematically transform the time-dependent observational data to the corresponding experimental data by leveraging the SCM. We then use the adjusted dataset for estimating HR using Cox PH Model. Our proposed approach focuses on causal effect of treatment on outcome to measure HR. We start with an observational study scenario and define all related assumptions. The schematic diagram for the proposed approach is shown in Figure 2.

4.1. Assumptions

We assume a simple observational study for a population, consisting of treatment X, confounding variable Z, and outcome in survival time T. In this scenario, treatment X is a dichotomous variable ($X = 1$ signifying treatment and $X = 0$ signifying control). Outcome T is the survival time from the beginning of the study and is a continuous variable in time units. Although the confounding variables can be a categorical or continuous variable, for simplicity, we assume the confounder Z to be a dichotomous variable. This observational study can be expressed as an SCM and with a graphical form G through causal directed acyclic graph (causal DAG) in Figure 3, where treatment, confounder, and outcome is expressed by the nodes X, Z and T respectively.
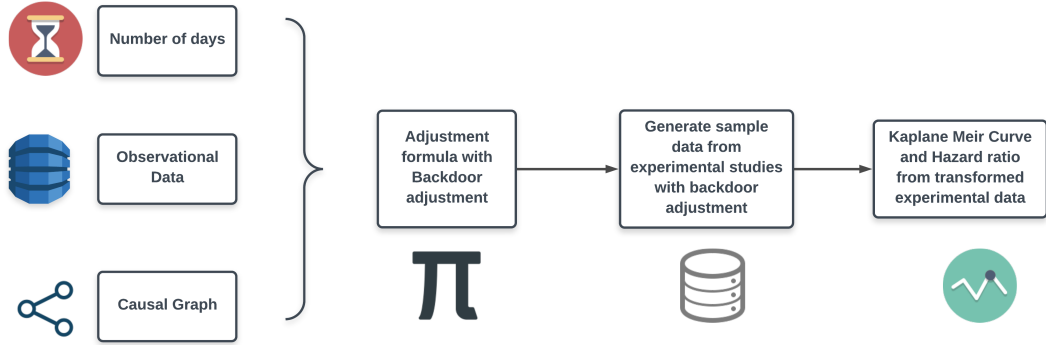
Figure 2: Schematic overview of the proposed approach. Observational data, corresponding causal diagram and the length of study is provided as input. The approach first use backdoor adjustment to create sample data from experimental study, and then compute the hazard ratio from the sampled experimental data.
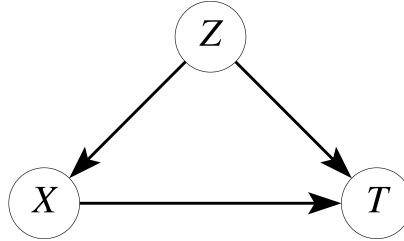

Figure 3: Simple observational study treatment X, outcome in survival-time T and single confounder Z, expressed using causal directed acyclic graph *(nodes are the variables, edges portray causal relationships between variables)*

From the definition of the SCM, we can express the underlying functions defining the causal relationships between variables by: $Z \leftarrow f_z(U_z), X \leftarrow f_x(Z, U_x), T \leftarrow f_t(Z, X, U_t, h_0(t))$. Here, $U = \{U_x, U_t, U_z\}$ is the set of exogenous variables, $V = \{Z, X, T\}$ is the set of endogenous variables, $f = \{f_z, f_x, f_t\}$ is the set of structural functions.

- $f_z(U_z)$ shows that confounder Z is independent of any other endogenous variables.

- $f_x(Z, U_x)$ expresses the dependency of X on Z. Since Z is parameter for both functions f_x and f_y, Z imposes a bias on the model $(P(X|Z = 0) \neq P(X|Z = 1))$, and the function f_x defines whether the bias is strong or weak.

- $f_t(Z, X, U_t, h_0(t))$ defines the effect of X and Z on the survival time T. This function also depends on the baseline hazard function $h_0(t, \mathbf{X})$ since this defines the rate of decline in survival.

We also assume to know the sample size of population n and a maximum length of survival time t_{max}.

4.2. Approach

4.2.1. TRANSFORMATION OF SINGLE STUDY TO MULTIPLE STUDIES

Experimental studies commonly have different study time-lengths, e.g., different number of days as the outcome endpoints (e.g., 30-day survival, 90-day survival, etc.). This variable is a dichotomous variable and describes a patients' status of survival at the end of the study. While analyzing a study similar to these, we do not take into account survival at each day, or survival after end-of-trial day, since we do not have the opportunity to do so. In our problem definition, we only have the survival time of individuals; however there is no defined end time for the trial. From the individual survival time, We can easily get the i-th day survival of every individual in the dataset, i being the number of days from the beginning of the study. We use *days* as smallest unit of time, since we assume the dataset reports survival in units of days. However, it could be any other units of times (*e.g. minutes, or weeks*) depending on the problem domain and dataset.

Since our observational study has a maximum survival time of all individuals t_{max}, we assume we calculate the variables Y_i, signifying the i-th day survival, i ranging from 0 to t_{max}. Conversion of continuous variable T describing survival time into multiple variables Y_i, each describing survival at the i-th day, essentially breaks down the single observational study into t_{max} number of observational studies with variables X, Z and Y_i, each of which is now a dichotomous variable.

Through the transformation, from a single SCM G, we end up with n different SCMs, each with the same treatment X and the confounder Z, but different outcome (survival at i-th day). Note that, in our assumption, the causal graph is time-invariant, i.e., the functional relationship between the variables does not change over time. This conversion is represented by n different SCMs (Transformed graphs A, Figure 4 (a)), where $n \geq t_{max}$.

An important point to note here is that, the single confounder Z and treatment X from the original observational study is not being transformed, only the outcome is distributed into multiple variables. In other words, we assume a point intervention and the confounding variables are invariant in time. And since we are transforming from a single trial to multiple trials, the outcomes Y_is of these separate trials are not conditionally dependent on each other (*e.g. a RCT with 30-day survival as outcome does not analyze about whether any patient died at 20th day or 29th day.*).

However, in extracting information from obs. data, there is dependency between them. Specifically, Y_i has causal effect on Y_j (where $j > i$), since if Y_i is 0 (*e.g. patient died at i-th day*), all Y_j (where $j > i$) is 0 (*e.g. patient remains dead for all consecutive days*). Also, Y_i only has direct causal effect on Y_{i+1}, every other corresponding effect is mediated through. If X has causal effect on Y_i, it is mediated through Y_{i-1}. For example, $X \perp\!\!\!\perp Y_1 | Y_0$, in absence of any backdoor variables. Reasoning behind this assumption is that, without having any underlying effect of treatment on outcome at i-th day, subject is suddenly prone to hazard on $i + 1$-th day. For example, this is unlikely that, if a subject is advised a treatment (drug), the subject has no hazardous effect until 10-th day and suddenly finds a hazardous effect on 11-th day. It is possible that the subject does not show any symptom

on 10-th day, or we cannot measure the internal hazardous effect of the drug on 10-th day (due to lack of symptoms).

The relationship between Y_is is reflected through a single transformed SCM (Transformed graph B, Figure 4 (b)), where $n \geq t_{max}$. The similarities between transformed graphs A and transformed graph B is that they both have same Z and X, and the dissimilarities are:

1. Transformed graphs A portrays n different trials with different outcomes, whereas transformed graph B is a single trial.

2. For transformed graphs A, $Y_i \perp\!\!\!\perp Y_j$ (where $j \neq i$), however for transformed graph B, $Y_i \not\!\perp\!\!\!\perp Y_j$ (where $j \neq i$).

3. Since two causal DAGs are different, transformed graphs A and transformed graph B have two different equations for $P(Y_i|do(X))$.

For outcome Y_0 For outcome Y_1 For outcome Y_n

(a) Converted Causal DAGs with no dependency between Y_is

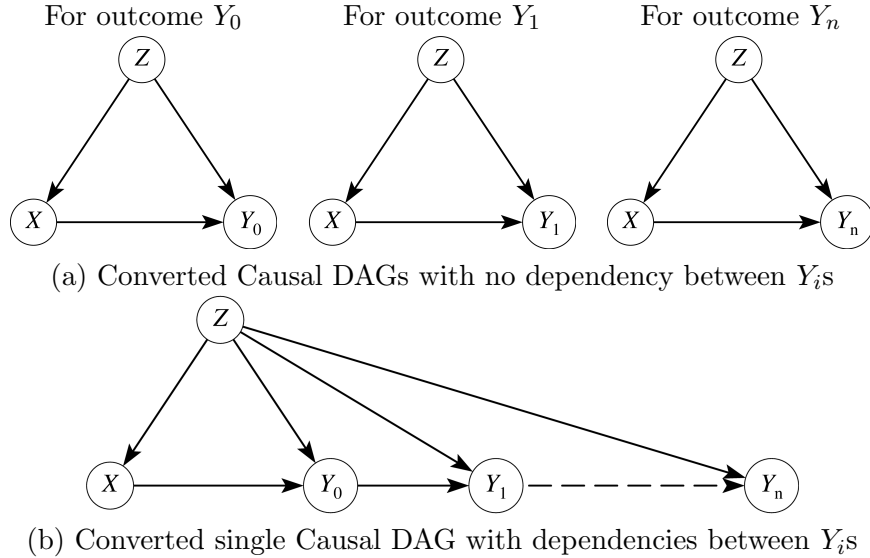(b) Converted single Causal DAG with dependencies between Y_is

Figure 4: Converted Causal DAGs with survival time converted to binary outcome of survival at different timepoints

In summary, we transform the single observational study into multiple different trials expressed through two different transformations (transformed graphs A and transformed graph B, Figure 4), each with the same treatment X and confounding Z, but with different survival time as the outcomes, as the death (or failure) increases over time. These outcomes are the status of survival (or death) at i-th day, where i is 0 to n ($n \geq t_{max}$).

4.2.2. Generation of Survival Curve

Applying Backdoor adjustment formula in transformed graphs A, the causal effect of X on Y_i (for all n causal graphs) is:

$$P(Y_i|do(X)) = \sum_Z P(Y_i|X,Z)P(Z)$$

In transformed graph B, the causal effect of X on Y_i is:

$$P(Y_i|do(X)) = \sum_{Z,Y_{i-1},...Y_1,Y_0} \left(\prod_{k=0}^n P(Y_k|Y_{k-1},...,Y_0,X,Z) \right) \cdot P(Z)$$

Since $P(A|B,C)P(B|C) = P(A,B|C)$ (using rules of conditional probabilities), we can reduce this equation to,

$$P(Y_i|do(X)) = \sum_{Z,Y_{i-1},...Y_1,Y_0} P(Y_i,Y_{i-1},...,Y_0|X,Z)P(Z)$$

Finally, for $j <= i$, $P(Y_j = 1|Y_i = 1) = 1$ (e.g. if a person is alive at 30th day, he has been alive for the last 29 days as well), $P(Y_i = 1, Y_{i-1} = 1) = P(Y_{i-1} = 1|Y_i = 1)P(Y_i = 1) = P(Y_i = 1)$, which reduces our equation down to the same as that of transformed graphs A:

$$P(Y_i = 1|do(X)) = \sum_Z P(Y_i = 1|X,Z)P(Z)$$

This signifies whether we use transformed graphs A or transformed graph B, we end up with same adjustment formula.

For each of the newly transformed causal DAGs, we can now adjust for the confounder using the backdoor adjustment formula. We calculate adjusted probabilities P_{adj} and thus adjusted counts C_{adj} for each of the n causal graphs. Using the values of P_{adj}, we generate survival curve with Kaplan Meier fitter.

4.2.3. Calculation of Hazard Ratio

Since we calculated C_{adj} for each of n causal graphs, we know number of adjusted individuals alive at each unit (day) of time. This helps us build back the adjusted survival time T_{adj} for individuals, as it was in the original dataset. The newly calculated survival time T_{adj} is adjusted for the confounding bias, as if they were sampled from an RCT. We measure the HR using Cox PH model with the adjusted survival time T_{adj} as outcome.

4.2.4. Algorithm

Algorithm 1 generates adjusted Kaplan Meier curve as well as the HR from Cox PH model on the adjusted dataset. The input for the algorithm is the dataset, specifically, confounder Z, treatment X, survival time T, and event status S. In the algorithm, variables in uppercase letters signify vectors, and variables in lowercase signify single variables. Internal procedures *convert_single_to_multiple_trials* (Algorithm 2) are shown separately.

Algorithm 1 Causally Formulated Hazard Ratio Estimation

Input: Z, X, T, S

Output: $survival\_curve, HR_{drug}$

 1: global $n \leftarrow length(T)$
 2: global $t_{max} \leftarrow max(T)$
 3: $Y_i \leftarrow convert\_single\_to\_multiple\_trials(T, S)$
 4: **for** $i \leftarrow 0$ to t_{max} **do**
 5:    $adj\_p_i \leftarrow \sum_Z P(Y_i = 1 \mid X, Z)P(Z)$
 6:    $adj\_c_i \leftarrow adj\_p_i * count(X)$
 7: **end for**
 8: $survival\_curve \leftarrow plot(time, cumulative(adj\_p_i))$
 9: $adj\_X, adj\_T \leftarrow convert\_multiple\_to\_single\_trial(adj\_c_i)$
10: $model \leftarrow cox\_ph\_model(adj\_X, adj\_T)$
11: $HR_{drug} \leftarrow exp(model.\beta_{drug})$
12: **return** $survival\_curve, HR_{drug}$

Algorithm 2 Conversion of single trial to multiple trials

Input: T, S

Output: Y_i

 1: **for** $i \leftarrow 0$ to t_{max} **do**
 2:    **for** $j \leftarrow 0$ to n **do**
 3:       $Y_i[j] \leftarrow (T[j] <= i)$ **and** $(S[j] = 1)$? $0 : 1$
 4:    **end for**
 5: **end for**
 6: **return** Y_i

---

5. Experiments and Applications

We evaluated the proposed approach for computing HR and visualizing survival curves with an experimental and observational dataset: (1) a synthetic dataset derived from a linear acyclic model with Gaussian noise; (2) a real-world dataset on disease-free survival in patients with Ewing's Sarcoma (Makuch, 1982). The rationale to consider these two datasets are: (1) both of the underlying causal model has a backdoor path through confounders, and, (2) both these datasets have treatment and control group that satisfy the proportionality hazarads assumption.

5.1. Experimental Data

We simulate an observational study with $n = 200$ patients. A subgroup of the patients received a treatment $(X = 1)$, and the remaining patients did not $(X = 0)$. We generate data on survival time T (in days) defined as the outcome. The treatment assignment is confounded by sex (e.g. $Z = 1$ for female, $Z = 0$ for others). The scenario has a causal model as depicted in Figure 3.

For the simulation, we generated outcome variable survival-time through defining a baseline hazard function. We defined survival time to be exponentially varying with time,
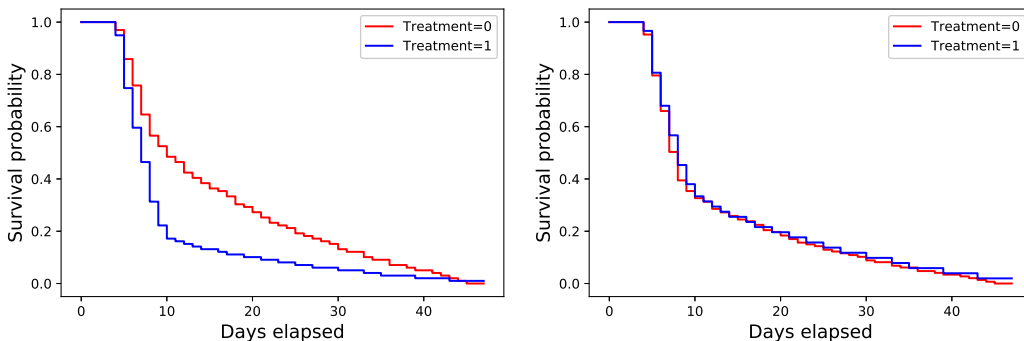
Figure 5: Unadjusted survival curve for simulated data *(left)* and, survival curve generated after applying proposed approach *(right)*. Figure on left shows significant difference in survival curve between treatment and control group. The treatment population ($X = 1$) seems to be more prone to hazard compared to the control population ($X = 0$). Figure on right shows adjusted survival curve to be overlapping, signifying no significant difference in hazard rate in both the treatment and the control population.

in the form of: $T \leftarrow a.exp((b+cZ+dX+eZX)*i)+E$, with Z being confounder, X being treatment, E being the noise/error and i being the index of patient. The other parameters were set to $a = 5, b = 0.025, c = 0.005, d = -0.015, e = 0.075, E = U(-0.5, 0.5)$, they were selected such that the HR remains close to 1, however injection of bias through Z portrays different outcome in survival curve.

We simulate the study with a strong biased effect from confounder Z. We define the strength of bias by an imbalance of conditional probabilities in each stratum of Z through the function $f_x(Z, U_x)$. For the defined strong bias case, $P(X = 1|Z = 0) = 0.75$ and $P(X = 1|Z = 1) = 0.25$. It translates to, if $Z = 0$ stands for females in this trial, 75% received the drug, whereas, in $Z = 1$ or others, only 25% received the drug. In a randomized controlled trial, under a no-confounding-bias scenario, we should have $P(X = 1|Z = 0) = P(X = 1|Z = 1) = 0.5$.

After we generate the experimental data, we applied Algorithm 1. We compared the existing approach of survival curve and survival curve from the adjusted dataset side-by-side in Figure 5.

Table 1 presents the HR found in the fitted Cox PH model in three different processes:

1. using only the treatment and outcome from the original dataset,

2. using data of all covariates (treatment, outcome and confounder) from the original dataset, and

3. using only the treatment and outcome from the adjusted dataset following our proposed approach.

The first approach represents scenarios where: 1) we ignore confounding, assuming it does not impact the treatment, or, 2) we do not possess data on the confounding variable
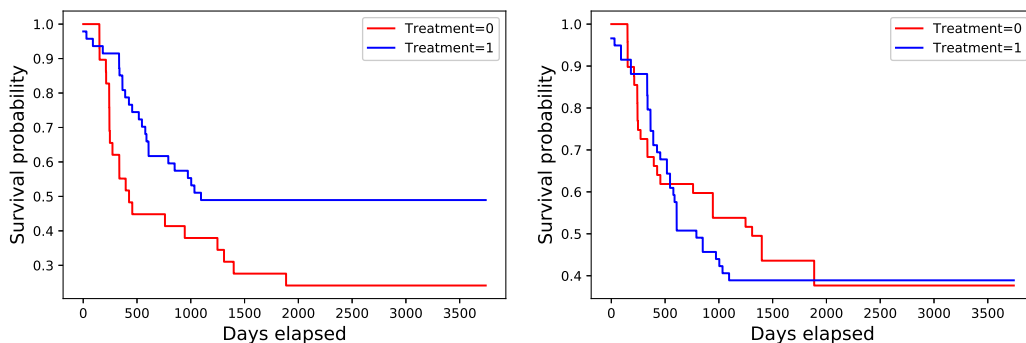
14

Figure 6: Unadjusted survival curve for Ewing dataset *(left)* and, survival curve generated after applying proposed approach *(right)*. Figure on left presents treatment group ($X = 1$) to be less hazardous than control group ($X = 0$). Figure on right is the adjusted survival curve with mostly overlapping survival curves of two groups, although treatment group ($X = 1$) seems slightly more prone to hazards.

(unmeasured confounding). This approach, however, results in an incorrect approximation of the HR. The second approach represents the existing approach to calculate HR. The third one shows our approach, and it eliminates the need for using confounding in model fitting since we are already adjusting for that.

Here, in Figure 5, the difference in unadjusted survival curve is similar to fitting Cox PH model with only X and T, leaving out Z, as found following the first approach generating HR=1.66. On the other hand, the overlapping adjusted survival curve is validated by calculated HR, following both the existing approach with the Cox PH model (HR=0.8) and our algorithm (HR=1.0).

5.2. Ewing's Sarcoma Data

We also applied the proposed method to a real-world dataset of patients with Ewing's Sarcoma (Makuch, 1982). The dataset was selected based on its survival data and known

	Existing model		Proposed model
	Observational data excluding confounding variable (biased estimate)	Observational data including confounding variable (traditional approach)	Transformed and adjusted data
Hazard Ratio	**1.66**	**0.80**	**1.00**
(95% Confidence Interval)	*(1.25-2.20)*	*(0.57-1.12)*	*(0.76-1.33)*

Table 1: Hazard Ratio for simulated dataset, calculated using existing model and our proposed approach. 1st column reports a biased estimate of HR, by using only treatment and outcome (excluding confounder) in Cox PH model. The second column reports a standard estimate of HR, by including all known variables (including confounder). The third column reports HR calculated in our proposed approach, using only treatment and outcome (excluding confounder).

15

causal DAG consisting of confounders. The dataset consists of a total of 76 Ewing's sarcoma patients with disease-free survival days as the outcome. 47 of the patients received a novel treatment (S4), and 29 received (one of) three (S1—S3) standard treatments.

The level of Serum lactic acid dehydrogenase (LDH) acted as the confounder, since high LDH levels indicated a lesser likelihood of treatment assignment along with an impact on survival time. In our analysis, we marked patients receiving S4 as the treatment group ($X = 1$) and patients receiving S1-S3 as the control group ($X = 0$). We applied our algorithm on this data set and the survival curve with the existing approach. Results of our algorithm is shown in Figure 6. The adjusted survival curve shows similar results, as demonstrated in Makuch et al. (Makuch, 1982). We also present the calculated HR following the three processes described in the earlier subsection. In Table 2, the HR calculated by our approach (HR=1.04) differs from the HR calculated in the traditional way (HR=1.12), presenting the drug to be a little less hazardous. However, the 95% confidence interval for both of these coincide, signifying that the true value lies within this range.

6. Discussion and Conclusion

We propose a novel method to estimate the HR using the Cox PH Model through the transformation of observational data to corresponding experimental data leveraging an underlying SCM. The transformed data are mathematically guaranteed to be adjusted for the confounding bias with the assumption that the SCM represents the data generating mechanism. Our woPrevious approaches under RCM that estimate the survival curve use the ignorability assumption, and will not work when the variables selected do not satisfy the backdoor criterion. Ignorability assumption states that, distribution of the potential outcomes $(Y(0), Y(1))$ is independent of the treatment variable by randomly assigning treatment: $\{Y(0), Y(1)\} \perp\!\!\!\perp X$. An extension of the idea, conditional ignorability states, distribution of the potential outcomes $(Y(0), Y(1))$ is independent of the treatment variable (X),

	Existing model		Proposed model
	Observational data excluding confounding variable (biased estimate)	Observational data including confounding variable (traditional approach)	Transformed and adjusted data
Hazard Ratio	**0.53**	**1.12**	**1.04**
(95% Confidence Interval)	*(0.30-0.96)*	*(0.59-2.11)*	*(0.57-1.87)*

Table 2: Hazard Ratio for Ewing dataset, calculated using existing model and our proposed approach. The first column reports a biased estimate of the HR based on only the treatment and outcome (excluding confounder) in Cox PH model. The second column reports a standard estimate of the HR that includes all known variables (including the confounder(s)). The third column reports the HR calculated in our proposed approach, using only the treatment and outcome (excluding the confounder). The reason for getting an accurate estimate of the HR even when excluding the confounder is because we adjusted the dataset beforehand using a minimum set of confounders from the SCM, thus focusing on the true causal effect of treatment on outcome.
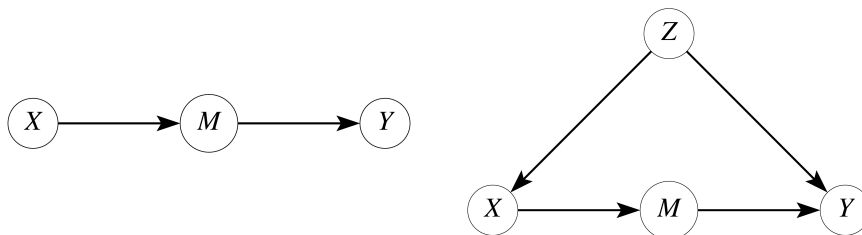
Figure 7: Two example graphs where the backdoor adjustment will produce different results compared to an approach based on the ignorability assumption. In the left hand side, X is the treatment, Y is the mediator, and M is a mediator. For the second graph, Z acts as a confounder as well. The left hand side is the example of a mediator and the right hand side graph is known as the front-door setting.

conditional on the covariates (Z): $\{Y(0), Y(1)\} \perp\!\!\!\perp X|Z$. Using conditional ignorability for adjustment on covariates allowed researchers to draw inferences from observational studies as well; however, adjusting all covariates irrespective of their causal relationship with treatment and outcome can contribute more bias to the model and incorrect estimation of causal effects.

We present two scenarios as example in Figure 7. In the first scenario (Figure 7, left), we show an SCM with treatment X and outcome Y with a third covariate M. Here M acts as a mediator in between X and Y, thus the backdoor adjustment gives a null set, meaning no adjustment is needed. The do-calculus formula would be: $P(Y|do(X)) = P(Y|X)$. Adjusting on M based on conditional ignorability will produce an incorrect estimation of causal effect. In the second scenario (Figure 7, right), we discuss a setting called front-door adjustment where we identify the variables to be adjusted with two applications of backdoor adjustment Pearl (2009). In an SCM with a mediator (shown in Figure 7 (right)), the covariate M does not satisfy the backdoor criterion and acts as a mediator between treatment X and outcome T. Thus, adjusting with M irrespective of its role as mediator will produce a biased estimate of the HR. The accurate backdoor adjustment formula (with M as mediator) is $P(Y|do(X)) = \sum_M P(M|X) \sum_X P(Y|X,M)P(X)$. However, adjustment by assuming M (and Z) as confounder gives an incorrect adjustment formula: $P(Y|do(X)) = \sum_{M,Z} P(Y|M,X,Z)P(Z)P(M)$. Our approach (with backdoor criterion) can correctly identify the variables to be adjusted for estimating HR using SCM and do-calculus.

Both the survival curve and the HR help to build a strong interpretation of the survival analysis of an experiment. The HR is most frequently reported as it summarizes the overall effect of treatment. However, survival curve encodes information on changes in survival over time (Hernán, 2010), which, in certain cases gives us better insight. Our proposed method is capable of generating both the survival curve and the HR, along with proper backdoor adjustment based on the underlying SCM. The HR calculated from the adjusted dataset requires only the treatment and outcome variables, and thus relies on direct causal relationships of treatment and outcome. For this purpose, we assume knowledge of the true causal model, an absence of unmeasured confounders, functional relationship in the SCM

being time-invariant, and, proportionality of the HR in the outcome. In reality, defining the causal graph with SCM, that is, causal structure learning, requires a principled approach. The development of statistical and computational algorithms for causal structure learning is an active research area (Heinze-Deml et al., 2018; Rottman and Keil, 2012), and, is not well-established in the current literature. We are currently working on a methodological framework to develop the causal graph with structure learning algorithm and domain expertise.

Acknowledgments

References

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455, 1996.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Stephen R Cole and Miguel A Hernán. Adjusted survival curves with inverse probability weights. *Computer methods and programs in biomedicine*, 75(1):45–49, 2004.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Vanessa Didelez and Nuala Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in medical research*, 16(4):309–330, 2007.

Ronald Aylmer Fisher et al. The design of experiments. *The design of experiments.*, (7th Ed), 1960.

Thomas R Frieden. Evidence for health decision making—beyond randomized, controlled trials. *New England Journal of Medicine*, 377(5):465–475, 2017.

Timothy D Girard, Matthew C Exline, Shannon S Carson, Catherine L Hough, Peter Rock, Michelle N Gong, Ivor S Douglas, Atul Malhotra, Robert L Owens, Daniel J Feinstein, et al. Haloperidol and ziprasidone for treatment of delirium in critical illness. *New England Journal of Medicine*, 379(26):2506–2516, 2018.

Jeremy A Greene and Scott H Podolsky. Reform, regulation, and pharmaceuticals—the kefauver–harris amendments at 50. *New England Journal of Medicine*, 367(16):1481–1483, 2012.

Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391, 2018.

Miguel A Hernán. The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.)*, 21(1): 13, 2010.

Miguel A Hernán and James M Robins. Observational studies analyzed like randomized trials and vice versa. In *Methods in Comparative Effectiveness Research*, pages 127–148. Chapman and Hall/CRC, 2017.

Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, pages 615–625, 2004.

Miguel A Hernán, Alvaro Alonso, Roger Logan, Francine Grodstein, Karin B Michels, Meir J Stampfer, Walter C Willett, JoAnn E Manson, and James M Robins. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology (Cambridge, Mass.)*, 19(6): 766, 2008.

Miguel Angel Hernán. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, 58(4):265–271, 2004.

Miguel Ángel Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, pages 561–570, 2000.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

David G Kleinbaum and Mitchel Klein. *Survival analysis*, volume 3. Springer, 2010.

David J Lederer, Scott C Bell, Richard D Branson, James D Chalmers, Rachel Marshall, David M Maslove, David E Ost, Naresh M Punjabi, Michael Schatz, Alan R Smyth, et al. Control of confounding and reporting of results in causal inference studies. guidance for authors from editors of respiratory, sleep, and critical care journals. *Annals of the American Thoracic Society*, 16(1):22–28, 2019.

Robert W Makuch. Adjusted survival curve estimation using covariates. *Journal of chronic diseases*, 35(6):437–443, 1982.

AD Nichol, M Bailey, DJ Cooper, On behalf of the POLAR, et al. Challenging issues in randomised controlled trials. *Injury*, 41:S20–S23, 2010.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

James M Robins. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pages 69–117. Springer, 1997.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Andrea Rotnitzky and James M Robins. Inverse probability weighting in survival analysis. *Wiley StatsRef: Statistics Reference Online*, 2014.

Benjamin M Rottman and Frank C Keil. Causal structure learning over time: Observations and interventions. *Cognitive psychology*, 64(1-2):93–125, 2012.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

Glen A Satten and Somnath Datta. The kaplan–meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3):207–210, 2001.

Spotswood L Spruance, Julia E Reid, Michael Grace, and Matthew Samore. Hazard ratio in clinical trials. *Antimicrobial agents and chemotherapy*, 48(8):2787–2792, 2004.