A PSEUDO-VALUE APPROACH TO CAUSAL DEEP LEARNING OF SEMI-COMPETING RISKS *

Stephen Salerno, PhD
Public Health Science Division, Biostatistics
Fred Hutchinson Cancer Center
Seattle, WA
ssalerno@fredhutch.org

Yi Li, PhD
Biostatistics
University of Michigan
Ann Arbor, MI
yili@umich.edu

ABSTRACT

While mortality is often the main focus of cancer studies, non-fatal events, such as disease progression, can vitally impact patient outcomes. For example, recurrence after curative treatment is a crucial endpoint in lung cancer, affecting available second-line treatments and personalized care. Estimating the de-confounded effect of interventions on disease recurrence is a key aspect of assessing cancer treatments. However, semi-competing risks complicate causal inference when death prevents disease recurrence. Existing approaches for estimating causal quantities in semi-competing survival functions rely on complex objective functions with strong assumptions and are challenging to estimate accurately. To address these challenges, we propose a deep learning approach for estimating the causal effect of treatment on non-fatal outcomes in the presence of dependent censoring and complex covariate relationships. Our three-stage approach involves estimating the marginal survival function using an Archimedean copula representation, and a jackknife pseudo-value approach that estimates pseudo-survival probabilities at fixed time points. These pseudo-survival probabilities serve as target values for developing causal estimators that are consistent and do not rely on assumptions like proportional hazards across all time points. In the final stage, we employ a deep neural network to link pseudo-outcomes, the causal variable, and additional confounders. This enables us to estimate survival average causal effects through direct standardization. We evaluate our approach through numerical studies and apply it to the Boston Lung Cancer Study, specifically examining the effect of surgical tumor resection in patients with early-stage non-small cell lung cancer.

1 Introduction

Lung cancer remains the leading cause of cancer-related deaths in the United States, accounting for one in five cancer deaths [Siegel et al., 2023]. Significant progress has been made towards improving lung cancer prognosis, owing in part to better screening and advances in targeted therapies [Liang et al., 2020]. However, a patient's clinical course may be highly variable due to the complex genetic, environmental, and psycho-social risk factors which influence disease progression and survival [Steliga and Dresler, 2011]. Furthering our understanding on the efficacy of patient-specific treatments is crucial when considering individualized approaches to care [Vargas and Harris, 2016, Politi and Herbst, 2015]. As it is not always practical to conduct randomized controlled trials due to ethical or practical reasons, *causal inference* is a powerful tool for making statements about the etiology of an outcome based on changes in a *causal variable* of interest in the context of observational studies [Pearl, 2009, Hernán and Robins, 2010].

In the context of survival analysis, common causal estimands include the average risk difference (i.e., difference in survival probability between treatment groups) at a given point in time or the average difference in restricted mean life time [Chen and Tsiatis, 2001, Rava, 2021]. Two common approaches for causal inference are outcome modeling and direct standardization or inverse probability weighting [IPW; Richardson and Rotnitzky, 2014]. In outcome modeling, the G-formula' is used to extend a standard regression model (referred to as the 'Q-model') for the conditional expectation of the outcome given the treatment and confounding variables. For survival endpoints, a natural

* *Citation*: **Authors. Title. Pages.... DOI:000000/11111.**

choice is the Cox Q-model, where the G-formula is applied to the Cox proportional hazards regression model [Andersen et al., 2017]. The Cox model has a well-established statistical theory and offers a straightforward method for hypothesis testing and inference [Cox, 1972]. Additionally, several specialized approaches have been developed to target causal parameters with time-to-event outcomes [Stitelman et al., 2012, Keogh et al., 2021, Andersen et al., 2017].

More broadly, while mortality is often the primary endpoint when studying the effect of a particular treatment or exposure, non-fatal events may also impact illness trajectories and treatment decisions related to disease management Kim et al. [2012]. In the context of lung cancer, disease progression alter remaining available treatments, making lung cancer recurrence in patients who have undergone curative treatment an important endpoint [Zappa and Mousa, 2016, Fedor et al., 2013]. Thus, having a comprehensive understanding of a patient's event history, in particular, disease progression is important to inform clinical decision making. It is often of substantial interest to study the 'net' effect of an intervention or exposure on time to disease progression [Delgado and Guddati, 2021]. However, there are two challenges that hamper this analysis – how to evaluate causality in observational studies [Gianicolo et al., 2020] and how to account for the the *semi-competing* relationship between disease progression and mortality [Jazić et al., 2016].

The presence of *semi-competing risks* can complicate causal inference by introducing dependent censoring, where the occurrence of death, or a *fatal* event, precludes recurrence, a *non-fatal* event. As a non-fatal event (recurrence) is often a precursor to the fatal event, this leads to informative censoring, which can bias estimates of treatment effects [Jazić et al., 2016, Ghosh, 2012, Nevo and Gorfine, 2022]. Much of the literature on causal methods for semi-competing risks are developed under a *potential outcomes* framework, using *principal stratification* to estimate causal effects [Nevo and Gorfine, 2022, Comment et al., 2019, Huang, 2021, Xu et al., 2022]. Principal stratification is a causal inference technique for handling post-treatment covariates in which patients are grouped based on post-treatment variables and causal effects are computed within these strata. For example, if we consider evaluating the causal effect of treatment, Z, on time to remission by time t_1, a principal stratification strategy would be to compute the survival average causal effect (SACE) among those individuals who would have survived as a member of either treatment or control group by some later time, t_2 [Frangakis and Rubin, 2002]. Here, the interpretation of the survival average causal effect (SACE) is causal effect on remission among those individuals who would have survived as a member of either the treatment or control group until at least time t_2. However, in many contexts, it is unclear as to whether principal stratification is truly of scientific interest, or if it is used to avoid ill-defined counterfactual outcomes [Pearl, 2011]. Further, many approaches for semi-competing survival functions use complicated objective functions, which require strong assumptions and are difficult to estimate with fidelity. Alternatively, when the outcome of interest is time to a non-fatal event, rather than the joint outcome of the non-fatal event and death, causal methods under the paradigm of 'truncation by death' have been developed [Zhang and Rubin, 2003, Long and Hudgens, 2013, Tchetgen Tchetgen, 2014, Zehavi and Nevo, 2021]. These approaches require special techniques to accommodate the presence of censoring.

Another promising approach to causal inference in survival analysis is through the use of pseudo-outcomes [Andersen et al., 2017]. Here, the time-to-event outcome, which is subject to censoring, is replaced by a pseudo-survival probability, which represents a given individual's contribution to estimating the survival function of the study sample. This approach has several benefits. Firstly, using a discrete time survival approach avoids the need for common assumptions. Typical strategies involve the use of parametric families to characterize the distributions of the survival times, which may be too restrictive in practice. Or, in the case of the Cox Q-model, the partial likelihood is defined under the assumption that the hazard ratio between different levels of each covariate (treatment and confounders) is constant over time (i.e., the proportional hazards assumption). This assumption may not hold, especially as the number of covariates increases. Pseudo-values address this issue by casting the outcome as the survival probability at given time points, allowing for non-proportional hazards. Further, pseudo-value based approaches replace the potentially censored survival times by jackknife-imputed survival probabilities. In the absence of censoring, standard loss functions can be used for optimization, rather than custom-designed approaches, and standard causal inference techniques such as IPW or the G-formula are applicable.

Despite these advances, often, parametric and semi-parametric methods are limited in their ability to model complex relationships and interactions between covariates [Jordan and Mitchell, 2015]. Typically, these approaches assume a linear relationship between the log survival time or log hazard and covariates, and non-linear relationships or complex interactions must be modeled explicitly. As such, there has been a growing interest in applying machine learning to survival analysis, in order to improve the accuracy of models [Wang et al., 2019, Sonabend et al., 2021]. Machine learning techniques, such as decision trees, random forests, and deep neural networks offer flexible and powerful approach for modeling survival data [Salerno and Li, Preprint posted online May 5, 2022]. These methods can account for complex covariate relationships and can handle high-dimensional datasets with many features. Several studies have demonstrated the effectiveness of machine learning approaches for survival analysis, including applications in cancer prognosis [Zupan et al., 2000, Cui et al., 2020, Doppalapudi et al., 2021, Wu et al., 2021]. Furthermore, the integration of causal inference into machine learning approaches has shown great promise for estimating the causal effects of treatments on survival outcomes. Several studies have proposed machine learning approaches for causal inference in

survival analysis. For example, Hu et al. [2021b] proposed an accelerated failure time Bayesian additive regression trees framework for estimating the heterogeneous survival treatment effects of lung cancer screening approaches, while Stitelman et al. [2012] proposed a general implementation of the targeted maximum likelihood estimator (TMLE) for longitudinal data in the context of a survival endpoint. These studies and others demonstrate the potential of combining machine learning techniques with causal inference methods for survival analysis.

Many recent developments have been made towards applying deep learning approaches for estimation to survival analysis [Yao et al., 2017, Katzman et al., 2016, 2018, Ranganath et al., 2016]. However, while the potential effects of covariates are indeed estimated non-parametrically as outputs from neural network architectures in these settings, the construction of these loss function relies on an underlying Cox proportional hazards or Cox frailty model, which may carry strong assumptions, or the survival times themselves may be assumed to arise from a parametric family of distributions. In such cases, there is a disconnect between these likelihood-based loss functions and common deep learning algorithms Steingrimsson and Morrison [2020]. Further applying deep learning to non-fatal event data with presents several challenges, including the need to account for dependent censoring, which requires careful modeling of the joint distribution of the semi-competing risks [Salerno and Li, 2022].

To address these issues, we propose a deep learning approach for estimating the causal effect of a given treatment on a non-fatal outcome in the presence of dependent censoring and potentially complex covariate relationships. In particular, we propose a three-stage approach. In the first stage, we estimate the marginal survival function for the non-fatal event based on a Clayton copula representation of the joint survival function. Following recent works by Andersen et al. [2017], Zhao and Feng [2020], Sabathé et al. [2020] and Orenti et al. [2021], we propose using jackknife pseudo-values to estimate pseudo-survival probabilities at fixed time points in the second stage. This circumvents the need for complex loss functions in downstream modeling, as using pseudo-survival probabilities reduces the problem at hand to minimizing the binary cross-entropy loss function. This also allows us to study causal targets which do not impose common assumptions such as proportional hazards across all time points. Lastly, to do so, we relate our pseudo outcomes to our causal variable of interest and additional confounders in a deep neural network to estimate survival average causal effect estimates via direct standardization.

The rest of this article is as follows. In Section 2, we introduce our notation and concepts such as the Clayton copula, jackknife pseudo-values, deep learning, and our target estimand for causal inference before outlining our three-stage procedure and formulating our deep neural network. In Section 3, we provide a series of numerical studies to evaluate our proposed approach, and in Section 4, we apply our method to the Boston Lung Cancer Study, a large scale epidemiologic lung cancer cohort study. We conclude with a discussion of our current work and areas of future research.

2 Method

2.1 Notation

We consider two event types – a non-fatal event, such as disease recurrence, and a fatal event (i.e., death), and introduce the following notation. For a study consisting of n individuals, let T_{i1} and T_{i2} denote the times to the non-terminal and terminal events, respectively, for the ith individual; $i = 1, \ldots, n$. We observe Z_i, the causal variable of interest, and X_i, a p-vector of additional confounding variables. In the context of our data, Z_i is binary treatment indicator taking values $Z_i = 1$ if a patient underwent surgical resection and $Z_i = 0$ for other first-line treatment options. Further, X_i includes demographics, prevalent comorbidity conditions, or genetic variants for the ith subject. We assume $(T_{i1}, T_{i2}, Z_i, X_i)$ are i.i.d. copies of (T_1, T_2, Z, X).

2.2 Bivariate Survival Function and the Clayton Copula

As a preamble, we consider a homogeneous situation, i.e., without covariates. We assume T_1 and T_2 are absolute, continuous random variables taking on non-negative values. Denote the marginal survival functions for the non-terminal and terminal events by $S_1(t_1) = Pr(T_1 > t_1)$ and $S_2(t_2) = Pr(T_2 > t_1)$, respectively. Note that the distribution of T_1 is non-parametrically identifiable only when the non-fatal event always precedes the fatal event [Xu et al., 2010]. Otherwise, as is the case in most practical settings, we assume a model for the joint survival distribution, given by

$$S(t_1, t_2) = \Pr(T_1 > t_1, T_2 > t_2).$$

When the non-terminal and terminal events are positively correlated, it is natural to assume a Clayton copula model to express $S(t_1, t_2)$ as a functional of marginal survival functions, $S_1(t_1)$ and $S_1(t_1)$ [Clayton, 1978], where

$$S(t_1, t_2) = [S_1(t_1)^{-\theta} + S_2(t_2)^{-\theta} - 1]^{-1/\theta} \tag{1}$$

and the copula dependence parameter, $\theta \geq 0$, measures the strength of the relationship between the non-fatal and fatal event times. Since the nonparametric function of (t_1, t_2) is only identifiable on the upper wedge, $0 < T_1 \leq T_2$, we assume model 1 on this upper wedge as well. Because model (1) may not hold in the lower wedge, the usual relationship that $\theta/(\theta + 2) =$ Kendall's τ may not hold [Fine et al., 2001].

2.3 Calculation of Distribution of Non-Fatal Event Time

Under the Clayton copula model, Fine et al. [2001] show that the marginal survival function for the non-fatal event time is monotonic and estimable given the joint survival function in (1) and the marginal survival function for the fatal event. Specifically, for a fixed time point, t, the joint survival function corresponds to the survival function for the first instance of either event, $S_*(t)$, which is often termed the *progression-free survival probability* in cancer research. The marginal survival function for the non-terminal event is related to the progression-free survival probability and the survival function for the terminal event via

$$S_1(t) = [S_*(t)^{-\theta} - S_2(t)^{-\theta} + 1]^{-\frac{1}{\theta}}, \tag{2}$$

which constitutes the basis of estimating $S_1(t)$, as both $S_*(t)$ and $S_2(t)$ are estimable via the Kaplan-Meier method, because both the time to the terminal event and the time to either event are always observable. Moreover, several works have proposed estimates for θ, including the estimator given by Oakes [1989] and Fine et al. [2001]. In the setting where the marginal survival functions do not depend on covariates, We can estimate θ "ad hoc" via the Oakes [1989] and Fine et al. [2001] concordance measure, given by

$$\frac{\sum_{i<j} W\left(Y_{ij1}, Y_{ij2}\right) D_{ij} \Delta_{ij}}{\sum_{i<j} W\left(Y_{ij1}, Y_{ij2}\right) D_{ij}\left(1 - \Delta_{ij}\right)} - 1. \tag{3}$$

where, for $1 \leq i \neq j \leq n$, we denote by $T_{ij1} = \min\left(T_{i1}, T_{j1}\right)$, $T_{ij2} = \min\left(T_{i2}, T_{j2}\right)$, and $C_{ij} = \min\left(C_i, C_j\right)$, and define $Y_{ij1} = \min\left(T_{ij1}, T_{ij2}, C_{ij}\right)$ and $Y_{ij2} = \min\left(T_{ij2}, C_{ij}\right)$ as the observable event times for the (i, j) pair. Further, $\Delta_{ij} = I\left[\left(T_{i1} - T_{j1}\right)\left(T_{i2} - T_{j2}\right) > 0\right]$ and $D_{ij} = I\left(T_{ij1} < T_{ij2} < C_{ij}\right)$, such that Δ_{ij} is estimable only when $D_{ij} = 1$. In contrast to the estimator of θ proposed in Fine et al. [2001], we make a modification in (3) by subtracting 1. This is because the definition of θ in our formulation (2) corresponds to $\theta + 1$ in Fine et al. [2001]. Lastly, let $Y_{i1} = \min(T_{i1}, T_{i2}, C_i)$ and $Y_{i2} = \min(T_{i2}, C_i)$ denote the observable event times for a given individual. The weight function, $W_{a,b}(y_1, y_2)$, is defined as

$$W_{a,b}^{-1}(y_1, y_2) = \frac{1}{n} \sum_{i=1}^{n} I\left\{Y_{i1} \geq \min(a, y_1), Y_{i2} \geq \min(b, y_2)\right\},$$

where constants a and b may be selected to dampen $W(\cdot)$ for large y_1 and y_2. Theoretically, Fine et al. [2001] show that $\hat{\theta}$ is a consistent estimator of θ, leading to the estimation of the non-fatal survival function in the absence of covariates.

2.4 Extension to the Distribution of Non-Fatal Event Time with Covariates

With covariates Z, X, the copula model (1) can be extended to

$$S(t_1, t_2 \mid Z, X) = \mathcal{C}_\theta[S_1(t_1 \mid Z, X), S_2(t_2 \mid Z, X)] = [S_1(t_1 \mid Z, X)^{-\theta} + S_2(t_2 \mid Z, X)^{-\theta} - 1]^{-1/\theta}, \tag{4}$$

where $S(t_1, t_2 \mid Z, X) = \Pr(T_1 > t_1, T_2 > t_2 \mid Z, X), S_1(t_1 \mid Z, X) = \Pr(T_1 > t_1 \mid Z, X)$ and $S_2(t_2 \mid Z, X) = \Pr(T_2 > t_2 \mid Z, X)$. Here, θ quantifies the correlation of T_1 and T_2 conditional on Z, X. Similarly, model (4) implies

$$S_1(t \mid Z, X) = [S_*(t \mid Z, X)^{-\theta} - S_2(t \mid Z, X)^{-\theta} + 1]^{-\frac{1}{\theta}},$$

which is the basis of estimating $S_1(t \mid Z, X)$. However, in this case, the estimator (3) of θ may not work as it was designed for a homogeneous population without considering covariates. Our idea is to extend estimator (3) by indirectly conditioning on Z, X. This is to mitigate information leakage, as we carry forward our estimate of θ to our downstream model. In particular, we propose to estimate $\hat{\theta}$ locally, in a neighborhood around (Z_i, X_i), by focusing on the nearest k neighbors to subject i, using the Euclidean distance between covariates. We run through all the subjects and average these estimates to achieve an overall estimate of θ. We term the procedure a 'leave-one-in' approach.

The rationale for this approach is that, by calculating θ among those observations with similar covariate distributions, the resulting estimate reflects a measure of concordance that is less sensitive to the impact of the covariates. More specifically, let X^* denote the matrix of covariates, including the treatment variable, Z, where each sample $X_i^* \in X^*$ is a $(p+1)$-dimensional vector. We consider the Euclidean distance, $D_{ii'}$, between X_i^* and $X_{i'}^*$ for $1 \leq i \neq i' \leq n$,

$$D_{ii'} = \|X_i^* - X_{i'}^*\|_2 = \left\{ \sum_{j=1}^{p+1} (x_{ij}^* - x_{i'j}^*)^2 \right\}^{1/2},$$

where x_{ij}^* and $x_{i'j}^*$ are the jth components of X_i^* and $X_{i'}^*$, respectively. Note, the Mahalanobis distance can also be used, and in our numerical experience, both distances work comparably. Then, for each individual, $i \in \{1, \ldots, n\}$, we identify the k nearest neighbors, among the n individuals, based on the their distances from this individual and denote them by $\mathcal{N}(i,k)$. We then estimate $\hat{\theta}$ based on subjects from $\mathcal{N}(i,k)$ via

$$\hat{\theta}^{(i)} = \frac{\sum_{j,l \in \mathcal{N}(i,k); j<l} W\left(Y_{jl1}, Y_{jl2}\right) D_{jl} \Delta_{jl}}{\sum_{j,l \mathcal{N}(i,k); j<l} W\left(Y_{jl1}, Y_{jl2}\right) D_{jl} \left(1 - \Delta_{jl}\right)} - 1.$$

Here, j and l index individuals in $\mathcal{N}(i,k)$. An overall estimate of θ is then given by

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}^{(i)}.$$

The number of neighbors, k, is chosen such that $|\hat{\theta}_k - \hat{\theta}_{k-1}| < \epsilon$, where $\hat{\theta}_k$ and $\hat{\theta}_{k-1}$ are the θ estimates with k and $k-1$ neighbors, respectively, and ϵ is a pre-specified tolerance level, say 0.01. See Figure 1 for an illustrative calculation over 50 generated datasets in which we vary the number of neighbors from 1 to 100 (black line = average value, grey ribbon = standard deviation), corresponding to Setting 2 in Section 3.

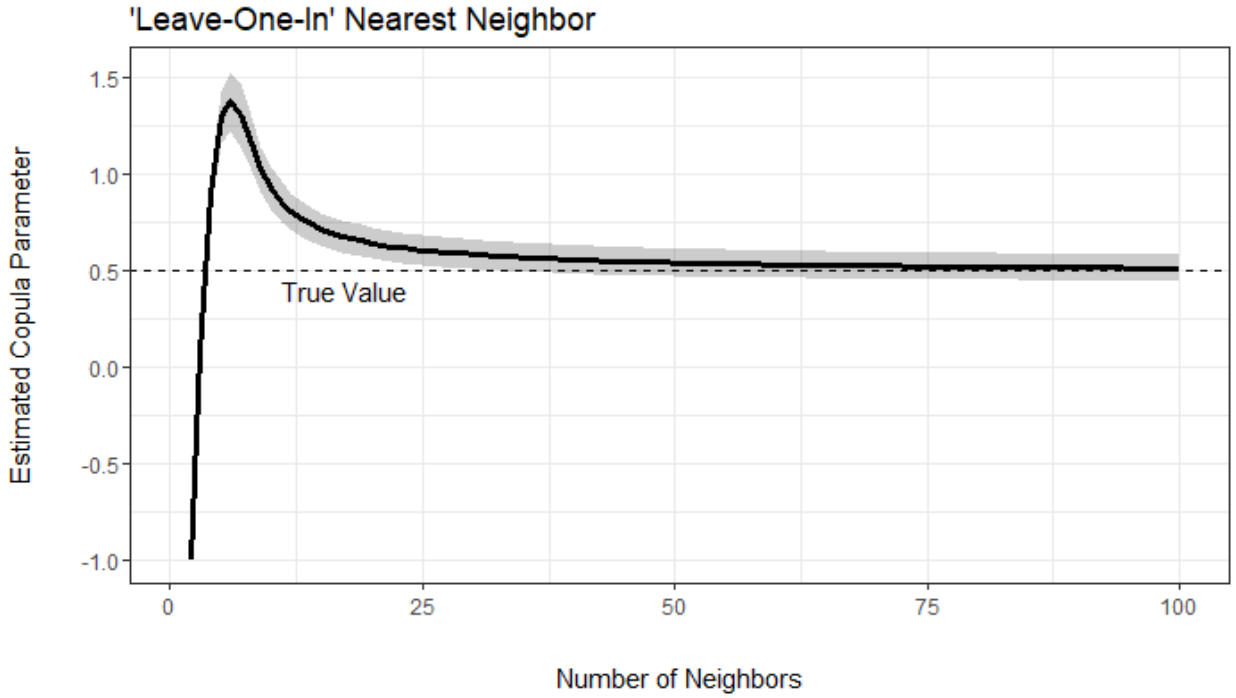


Figure 1: Example calculation across 50 simulated datasets with correlated covariates

2.5 Potential Outcomes Framework for Causal Inference

Under a potential outcomes framework, T_{i1}^z denotes the potential time to recurrence that would occur had $Z_i = z \in \{0, 1\}$ for the ith individual. Causal inference infers the 'true' effect of an intervention on time to disease recurrence by comparing T_{i1}^1 versus T_{i1}^0 [Pearl, 2009]. Before proceeding we make several common assumptions:

1. **Consistency**: $\exists \{T_{i1}^1, T_{i1}^0\}$ s.t. $T_{i1} = T_{i1}^{Z_i}$ almost surely. In other words, an individual's potential outcome under their assigned treatment group is the outcome that will actually be observed.
2. **No Interference**: T_{i1}^z is unaffected by the value of z for another subject, j.
3. **Positivity**: $Z_i \in \{0, 1\}\ \forall X_i$, or the assumption that every individual has a non-zero probability of being assigned to either treatment group.
4. **Exchangeability**: $T_{i1}^1, T_{i1}^0 \perp Z_i \mid X_i$, i.e., 'no unmeasured confounding.'

Together, the assumptions of *consistency* and no interference make up the stable unit treatment value assumption (SUTVA), which underpins the existence of the counterfactual outcomes, T^z. We require positivity to ensure that our observed data functionals, denoted below by $S_{i1}(t \mid X_i, Z_i = z)$, are well defined. Lastly, the exchangeability assumption allows us to represent the conditional distribution of the unobserved potential outcome using that of the observed potential outcome. We are interested in the average causal effect of Z_i on the time to recurrence, T_{i1}. With the assumptions above, a common causal quantity of interest given the counterfactual potential outcomes is the average treatment effect (ATE), or the expected difference in potential outcomes over all individuals in the study. We can consider the average causal difference in the risk of recurrence at time t as

$$\mathbb{E}[I(T_{i1}^1 > t)] - \mathbb{E}[I(T_{i1}^0 > t)], \tag{5}$$

For Equation (5), note that $\mathbb{E}[I(T_{i1} \leq t)] = 1 - S_1(t)$. Thus, given a consistent estimator of $S_1(t)$, $\hat{S}_1(t)$, we can estimate the ATE by training a deep neural network for $S_{i1}(t \mid X_i, Z_i)$ and predicting the potential outcomes $\hat{S}_{i1}(t \mid X_i, z); z \in \{0, 1\}$. An estimate of the ATE for the average causal risk difference is given by

$$\hat{\text{ATE}} = n^{-1} \sum_{i=1}^{n} \{\hat{S}_{i1}(t \mid X_i, 1) - \hat{S}_{i1}(t \mid X_i, 0)\}. \tag{6}$$

2.6 A Pseudo-Values Approach for Causal Estimation

Our goal is to construct a model to study the difference in risk of recurrence at a given point in time. As the efficacy of a given treatment may change over time, common approaches to causal survival analysis such as the Cox Q-model may impose certain structures across all time points, e.g., proportional hazards, that are not realistic. Pseudo-values provide an intuitive means of circumventing the proportional hazards assumption, while also replacing potentially incompletely observed outcomes with a real-valued function of our outcome for each individual [Andersen et al., 2017].

For any function, $f(t)$, jackknife pseudo-responses can be generated as $\hat{f}_i(t) = n\hat{f}(t) - (n-1)\hat{f}^{-i}(t)$, where $\hat{f}(t)$ is the overall estimate of $f(t)$ and $\hat{f}^{-i}(t)$ is an estimate omitting the ith subject. In our setting, consider J discrete time points, indexed by $j = 1, \ldots, J$. The probability of no recurrence by time t_j is given by $S_1(t_j) = \Pr(T_{i1} > t_j)$. A pseudo-outcome for individual i at time point t_j can be constructed as

$$\hat{S}_{i1}(t_j) = n \times \hat{S}_1(t_j) - (n-1) \times \hat{S}_1^{-i}(t_j)$$

where $\hat{S}_1(t_j)$ and $\hat{S}_1^{-i}(t_j)$ are the overall estimate of $S_1(t_j)$ using all n subjects and the 'leave-one-out' estimate excluding the ith subject, respectively, based on (2). Intuitively, this estimator for $S_1(t_j)$ represents the contribution of the ith individual in estimating $\mathbb{E}[S_1(t_j)]$ in a sample of n subjects. Further, because we have a consistent estimate of $S_1(t)$, $\hat{S}_{i1}(t_j)$ is approximately independent of $S_{i'1}(t_j)$ for $i \neq i'$ as $n \to \infty$ and

$$\lim_{n \to \infty} E[\hat{S}_{i1}(t_j) \mid X_i, Z_i] = S_1(t_j \mid X_i, Z_i)$$

for any i [Ahn and Mendolia, 2014, Logan et al., 2011]. With these results, the pseudo-values, $\hat{S}_{i1}(t)$, can then be used as numeric responses, similar to a logistic model fit to $I\left(T_{i1} > t_j\right)$ if the data were fully observed. However, as

$I(T_i > t)$ is not observed for all subjects due to censoring, we must estimate the pseudo-responses for both the censored and uncensored individuals. We carry forward a design matrix of size $n \times (p + J)$, where p denotes the number of covariates in X and we include $J - 1$ dummy variables encoding time t_j.

2.7 Neural Network Architecture

Rather than parameterizing the effects of the covariates as a linear function, we instead formulate a deep learning approach for estimation. As shown previously, the benefit here is that we can better capture potential non-linear and higher-order dependencies between covariates that make up a complex risk profile for patients, including high-dimensional covariates [Bauer and Kohler, 2019, Poggio et al., 2017]. The pseudo-value approach facilitates direct estimation of the target quantity of interest, without needing to optimizing the joint likelihood of the survival times directly. This circumvents the need for complex loss functions as part of the neural network architecture. Our deep neural network (DNN) directly minimizes the binary cross-entropy loss between the pseudo-survival probabilities, $\hat{S}_{i1}(t_j)$, and the predicted survival probabilities from the neural network output, $\pi_i(t_j)$, such that

$$\text{Binary Cross Entropy Loss} = \frac{1}{n}\left\{\sum_{i=1}^{n} -\hat{S}_{i1}(t_j)\log[\pi_i(t_j)] - [1-\hat{S}_{i1}(t_j)]\log[1-\pi_i(t_j)]\right\}. \tag{7}$$

Our proposed DNN is referred to as an *S*-learner, as it consists of a ***single*** fully-connected feed-forward neural network with an input layer, L hidden layers with k_l neurons in the lth layer; $l = 1, \ldots, L$, and an output layer [Zhao and Hastie, 2021, Koch et al., 2021]. Hidden layers are connected via a non-linear activation function such as the rectified linear unit activation functions (ReLU; $\sigma_l(x) = \max(0, x)$), while the output layer's activation function is specified based on the target quantity. For example, as our target values are survival probabilities, a sigmoidal activation function ($\sigma_l(x) = \{1 + e^{-x}\}^{-1}$) is used for the final layer to constrain the output probabilities between 0 and 1. Estimation is based on an L-fold composite function

$$F_L(\cdot) = f_L \circ f_{L-1} \circ \cdots \circ f_1(\cdot) \text{ where } (g \circ f)(\cdot) = g(f(\cdot)),$$

$$f_l(x) = \sigma_l\left(\mathbf{W}_l x + \mathrm{b}_l\right) \in \mathbb{R}^{k_{l+1}},$$

where σ_l is an activation function, $\mathbf{W}_l$ are weights, and b_L are biases. Our network output is optimized under (7), which has a faster convergence rate than the traditional mean squared error due to its steeper gradient when the predicted output is far from the true output. Our final layer outputs a representation of the data, Ψ, which is used to then predict the counterfactual outcomes for each individual, $\hat{S}_{i1}(t \mid X_i, z); z \in \{0, 1\}$, before calculating (6).

Hyperparameters needed to fully specify the neural network architecture include the number of hidden layers and number of nodes per hidden layer, the dropout fraction, and learning rate. In practice, these quantities are optimized over a Cartesian grid search based on predictive performance. We implement our approach with the `R` interface for Keras, using the deep learning library TensorFlow as the backend [Allaire and Chollet, 2022, Allaire and Tang, 2022]. Software to implement this method can be found at `https://github.com/salernos/pseudoSCR`

3 Simulations

3.1 Data Generation

We next performed a series of simulations to assess the accuracy of our proposed approach against standard methods. In particular we varied the sample size, copula dependence parameter, censoring rates, and covariate-dependent risk functions in a fully factorial design. We considered two cases for the sample sizes, letting $n = 500$ or $n = 1,000$. Further, we let the copula dependence parameter, θ, equal 0.5, or 2. Noting that Kendall's $\tau \approx \theta/(\theta + 2)$, these setting correspond to approximate Kendall's τ values of 0.2 or 0.5, respectively. Dependent on each data generation model, we varied the parameters used to generate censoring times to achieve approximate censoring rates of 0% or 50%. Lastly, we considered two different generative model settings, described further below.

We considered two generative models of varying complexity. In the first setting, we simulated data from a proportional hazards model with a risk function that is linear in terms of the covariates, facilitating a fair comparison between the competing methods. In the second setting, we again simulated the data from a proportional hazards model, but we introduced a non-linear risk function through the use of higher order terms and correlated covariates.

Setting 1: Linear Risk Function. We first generated the data following the simulation scheme proposed in Peng and Fine [2007], Hsieh and Huang [2012], and Orenti et al. [2021]. Specifically, we generated non-fatal (T_{i1}) and fatal (T_{i2}) event times from marginal models specified by

$$\log(T_{i1}/3) = -(\beta_1 Z_i + \beta_1 X_{i1} + \beta_1 X_{i2}) + \varepsilon_{i1}$$
$$\log(T_{i2}/3) = -(\beta_2 Z_i + \beta_2 X_{i1} + \beta_2 X_{i2}) + \varepsilon_{i2},$$

where Z_i is a Bernoulli random variable with a success probability of 0.5, X_{i1} and X_{i2} are independent truncated normal random variables with mean 1, variance 0.5, and truncation bounds of [0, 2], and $(\varepsilon_{i1}, \varepsilon_{i2})$ are correlated random errors. To induce dependence between the event times, we simulated ε_{i1} and ε_{i2} from the Clayton copula model,

$$\left[\Pr\left(\varepsilon_{i1} > t_1\right)^{-\theta} + \Pr\left(\varepsilon_{i2} > t_2\right)^{-\theta} - 1\right]^{-\frac{1}{\theta}},$$

where ε_{i1} and ε_{i2} follow the extreme value distribution, i.e., $\Pr\left(\varepsilon_{i1} > t_1\right) = \exp\{-\exp(t_1)\}$ and $\Pr\left(\varepsilon_{i2} > t_2\right) = \exp\{-\exp(t_2)\}$ [Rotolo et al., 2013]. See Appendix A for additional details.

Setting 2: Non-Linear Risk Function, Correlated Covariates. In our second data generation scenario, we adopted a similar framework as described previously, but we have modified the covariate risk functions to include higher-order terms and correlations to understand the performance differences between our non-parametric approach and approaches which are misspecified when assuming a linear form with independent covariates. We generated three covariates, $\boldsymbol{X} = (X_1, X_2, X_3)'$, from a multivariate normal distribution with $\boldsymbol{X} \sim N_3(\mathbf{0}, \Sigma)$, where the covariance matrix, Σ, is AR(1) with elements $(\sigma_{ij}) = 0.5^{|i-j|}$. We then dichotomized $Z_i = \mathbb{I}(X_{i1} \geq 0)$ to be a binary covariate representing our causal variable of interest. We generated the event times, T_{i1} and T_{i2}, from marginal models specified by

$$\log(T_{i1}/3) = -(\beta_1 Z_i + \beta_1 X_{i1}^2 + \beta_1 X_{i2}^2) + \varepsilon_{i1}$$
$$\log(T_{i2}/3) = -(\beta_2 Z_i + \beta_2 X_{i1}^2 + \beta_2 X_{i2}^2) + \varepsilon_{i2},$$

Across all scenarios, we fixed $\beta_1 = 1$ and $\beta_2 = 0.2$. In settings where the event times may be censored, we generated independent censoring times, C_i, from a mixture of uniforms, where $C_i \sim \text{Unif}(0, 1)$ with probability 0.2 and from $\text{Unif}(1, 1.2)$ with probability 0.8, yielding an approximate censoring rate of 50%. For each combination of settings, we generated 50 independent datasets and calculated the average bias and mean squared error (MSE) for the estimated average treatment effect (ATE) for our proposed approach against a causal Q-model, which was fit using generalized estimating equations with a complementary log-log mean link, corresponding to the proportional hazards model [Orenti et al., 2021]. To calculate the pseudo-values, we first estimated the copula dependence parameter using the 'leave-one-in' approach described previously, applied to the entire sample of n observations. We carried forward the estimated $\hat{\theta}$ to calculate the pseudo- non-fatal survival probabilities at fixed time points $t = 0.2$, 0.4, 0.6, 0.8, and 1.0. For our method, we hypertuned our DNN parameters once per simulation setting and carried forward the best configuration of hyperparameters across all 50 datasets. Lastly, we randomly split each dataset into an 80% training set and a 20% testing set. We fit the respective models on the training set and calculated the ATE at $t = 1.0$ in the testing set.

3.2 Example Pseudo-Value Calculation on Simulated Data

First, we illustrate the calculation of the pseudo-values on simulated data. Table 1 gives examples of the estimated pseudo-values for two individuals from one simulated dataset under the first data generating mechanism. We selected an individual from each hypothetical treatment arm, where Individual 1 was simulated to have the control ($Z_i = 0$) and Individual 2 was simulated to have the treatment ($Z_i = 1$). Further, these individuals were chosen for illustration as Individual 1 experienced both the non-terminal and terminal events before one year of follow-up, whereas Individual 2 was administratively censored shortly after the one year mark. As shown, the estimated pseudo-recurrence probability for Individual 1 at time $t = 0.2$ was close to 1, as the individual did not experience the non-terminal event until time $t = 0.3991$. Subsequently, the pseudo-probabilities for times $t = 0.4$ to $t = 1.0$ are close to 0, as these time points are after the event occurred. In contrast, the pseudo-probabilities for Individual 2 are all approximately 1 across all time points, as this individual was censored after time $t = 1.0$. Lastly, due to the presence of censoring, the pseudo-probabilities are real-valued and not restricted to $\{0, 1\}$ [Andersen and Pohar Perme, 2010].

Table 1: Example pseudo-values for two individuals.

Observation		Simulated Outcomes				Treatment	Estimated
ID	t	Y_{i1}	D_{i1}	Y_{i2}	D_{i2}	Z_i	**Pseudo-Values**
1	0.2	0.3991	1	0.4054	1	0	1.0302
1	0.4	0.3991	1	0.4054	1	0	-0.3260
1	0.6	0.3991	1	0.4054	1	0	0.1765
1	0.8	0.3991	1	0.4054	1	0	0.0968
1	1.0	0.3991	1	0.4054	1	0	0.0496
2	0.2	1.0401	0	1.0401	0	1	1.0302
2	0.4	1.0401	0	1.0401	0	1	1.1761
2	0.6	1.0401	0	1.0401	0	1	1.3082
2	0.8	1.0401	0	1.0401	0	1	1.4430
2	1.0	1.0401	0	1.0401	0	1	1.5688

3.3 Simulation Results

Table 2 summarize the results of this simulation study. As shown, model performance was similar in the first data generation setting where the parametric Q-model is correctly specified, though the correct model is slightly less biased and more efficient. This is to be expected, as we are fitting the true model to the data, while the DNN represents a stochastic approximation of the true data generation function. In the second setting, however, the performance for our proposed approach is better, as the true covariate risk function contains correlated covariates and higher-order terms. While the degree of bias for the proposed approach remains fairly consistent with the first data generation setting, the bias increases for the parametric Q-model. We also note that for both methods, performance was typically better in settings with a larger sample size ($n = 1,000$ versus 500), a smaller degree of dependence between the event times ($\theta = 0.5$ versus 2.0), and when the data were fully observed versus censored, as expected.

3.4 Sensitivity Analysis

In a sensitivity analysis, we study the performance of the proposed approach against model misspecification. Specifically, we generate data from the marginal models described above, but to induce dependence between the simulated event times, we now generate the error terms from the Gumbel copula, rather than the assumed Clayton copula. The bivariate Gumbel copula is given by

$$\exp\left\{-\left[\log \Pr\left(\varepsilon_{i1} > t_1\right)^{\theta} + \log \Pr\left(\varepsilon_{i2} > t_2\right)^{\theta}\right]^{\frac{1}{\theta}}\right\}.$$

Like the Clayton copula, the Gumbel copula cannot have negative dependence, and it converges to the co-monotonicity copula as $\theta \to \infty$. However, as the Gumbel is the independence copula when $\theta = 1$, rather than $\theta = 0$, we consider only the simulation setting where $\theta = 2$ for this sensitivity analysis [Ruppert and Matteson, 2011]. As shown, we do incur bias if our model is misspecified for the data generating copula, with the ATE tending to be underestimated for the Gumbel copula. We further see that we have a higher mean squared error across all settings when the data are generated from the Gumbel copula, as compared to the Clayton copula. For additional details and full results, see Appendix A.

4 Boston Lung Cancer Study

The Boston Lung Cancer Study is a collaborative research effort between Dana-Farber Cancer Institute and Massachusetts General Hospital which focuses on improving the understanding and treatment of lung cancer, one of the leading causes of cancer-related deaths worldwide [Christiani, 2017].

4.1 Study Population

Among all participants in the Boston Lung Cancer Study (BLCS) cohort, 7,755 were initially eligible for inclusion in this analysis. Eligibility was defined as having a positive lung cancer diagnosis. Participants were ineligible if they were enrolled with esophageal cancer or other primary cancer, no cancer upon further study, or as a negative control in

Table 2: Average bias and mean squared error (MSE) for estimated vs. true ATE comparing our proposed method to the parametric Q-Model. Results are averaged over 50 independently generated datasets for each setting.

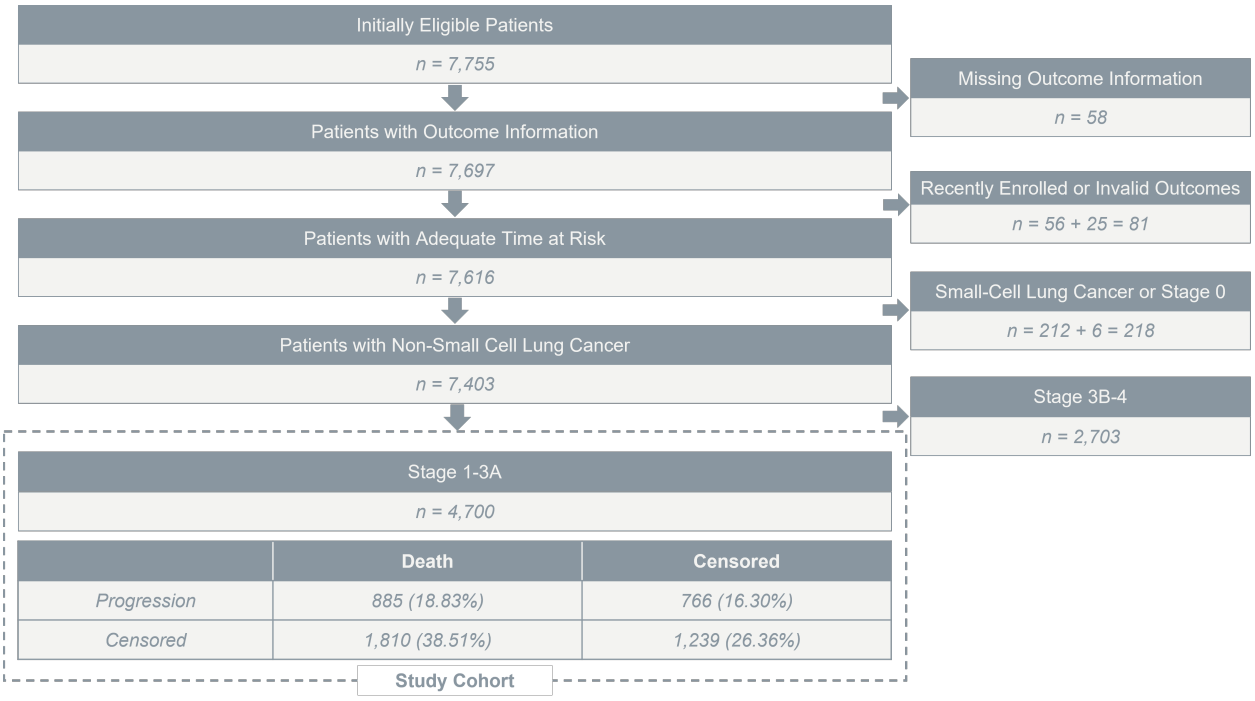
Simulation Settings				Bias		Mean Squared Error	
n	θ	τ	Censoring	Q-Model	Proposed	Q-Model	Proposed
Setting 1: Linear Risk Function							
500	0.5	0.2	50%	0.0025	0.0060	0.0020	0.0063
500	0.5	0.2	0%	0.0025	0.0045	0.0022	0.0042
500	2.0	0.5	50%	0.0025	0.0057	0.0022	0.0053
500	2.0	0.5	0%	0.0018	0.0069	0.0019	0.0011
1000	0.5	0.2	50%	0.0018	0.0025	0.0013	0.0028
1000	0.5	0.2	0%	0.0023	0.0035	0.0014	0.0028
1000	2.0	0.5	50%	0.0019	0.0048	0.0014	0.0037
1000	2.0	0.5	0%	0.0018	0.0030	0.0012	0.0021
Setting 2: Non-Linear Risk Function, Correlated Covariates							
500	0.5	0.2	50%	0.0483	0.0043	0.0076	0.0032
500	0.5	0.2	0%	0.0520	0.0030	0.0078	0.0031
500	2.0	0.5	50%	0.0444	-0.0083	0.0081	0.0045
500	2.0	0.5	0%	0.0476	-0.0030	0.0079	0.0046
1000	0.5	0.2	50%	0.0485	-0.0043	0.0036	0.0028
1000	0.5	0.2	0%	0.0518	-0.0034	0.0038	0.0024
1000	2.0	0.5	50%	0.0444	-0.0040	0.0046	0.0032
1000	2.0	0.5	0%	0.0475	-0.0035	0.0042	0.0033

the case of spouses, friends, or other participants. Among those 7,755 eligible patients, we identified 7,697 (99%) with the temporal information necessary to define their semi-competing outcomes, namely (1) date of primary diagnosis, (2) recurrence, progression, and/or death date where applicable, and (3) last follow-up date or non-progression date. We further removed 56 patients diagnosed in the past 6 months, 25 patients with negative survival times, 212 patients with small-cell lung cancer, and 6 patients with carcinoma *in situ* , i.e., stage 0 (Figure 2). As available treatment options are predicated on a patient's cancer stage, we considered two subgroups of patients – those who were diagnosed with stages 1-3a NSCLC (4,700; 63.5%) and those who were diagnosed with stages 3b-4 NSCLC (2,703; 36.5%). As stages 1-3a are widely considered to be operable, we focused on understanding the average treatment effect of first-line surgical resection on time-to-relapse among this subset of patients (Figure 2).

4.2 Patient Characteristics

Descriptive statistics for the study cohort are given in Table 3. As shown, median age among all patients with NSCLC was 66 years old [interquartile range (IQR): 59-74], with a majority of patients identifying as female (54%), White/Caucasian (92%) and non-Hispanic (87%). Further, the majority of study participants were former smokers (57%) with a median 40 pack-years of smoking (IQR: 16-53). Among all patients, the majority underwent surgical resection (4,444; 67%) as first-line treatment. However, stratifying by stage at diagnosis, we found that patients with earlier-stage diagnoses were slightly older (median, IQR age: 68, 61-74 years versus 64, 56-72 years), with a higher proportion being female (55% versus 50%) and White/Caucasian (93% versus 92%), and a lower proportion identifying as non-Hispanic (85% versus 90%). Social history differed between these two groups as well, with more former smokers (60% versus 53%) as compared to current smokers (25% versus 30%) in the earlier-stage group, though a higher median number of pack-years of smoking (40 versus 37 pack-years). Lastly, rates of testing for two common genetic variants, EGFR and KRAS, differed between these groups, with more patients (81% versus 76%) tested in the earlier-stage group. Among those tested, we observed a higher proportion of patients in the earlier-stage group with a KRAS mutation (30% versus 21%), though a higher proportion in the late-stage group with an EGFR mutation (18% versus 21%). We then carried forward our final analytic cohort of 4,700 patients diagnosed with non-small cell lung cancer (NSCLC), stages 1-3a. Disease recurrence was reported in 1,651 (35.13%) patients, with 885 (18.83%) patients experiencing recurrence followed by death and 1,810 (38.51%) patients who died prior to recurrence (Figure 2).

Figure 2: Flowchart of inclusion and exclusion criteria for the Boston Lung Cancer Study analytic sample and distributions of observed outcomes (progression and/or death).



4.3 Time-to-Recurrence and Estimated Pseudo-Values

In line with our proposed analytic framework, we first calculated the survival function for recurrence based on the joint survival function and the survival function for death under the assumed Clayton copula. We calculated this for the entire study sample, as well as stratified by patient sex (male versus female). The copula dependence parameter, θ, captures the strength of the relationship between progression and death, with larger values corresponding to a higher degree of dependence between these two events. We estimated the value of this parameter using our 'leave-one-in' modification to the extended concordance-based estimator proposed in Fine et al. (2001) [Fine et al., 2001]. Among all patients in our study, we estimated the dependence between progression and death to be 5.60, corresponding to an approximate Kendall's τ value of 0.737. This suggests a high degree of correlation between progression and death. Further stratified by patient sex, we estimated this dependence to be higher among females (5.93) than males (4.85), corresponding to approximate Kendall's τ values of 0.748 versus 0.708, respectively.

We then estimated the marginal time-to-recurrence distribution and jackknife pseudo-survival probabilities. We calculated pseudo-recurrence probabilities at one-year benchmarks from one- to five-years follow up. Figure 3 gives the distribution of the estimated pseudo-values for the probability of recurrence at each year of follow-up and stratified by first-line treatment. Values on the y-axis are standardized so that each bar represents the proportion of patients within each treatment group for the specified bin width. Note that the predicted survival probabilities are not strictly 0 or 1 in the presence of censoring, nor are the values confined to [0,1]. Instead, the pseudo probability is a real value which takes on an approximately bimodal distribution [Andersen and Pohar Perme, 2010, Zhao and Feng, 2020]. For each treatment group, the distribution shifts from 1 toward 0 in each successive year of follow-up. However, there is a higher relative proportion of patients who received other first-line treatments with lower survival probabilities than those patients who underwent surgical resection.

4.4 Risk Difference between First-Line Therapies

We carried forward these pseudo-outcomes to our S-learner, where we estimated the average causal difference in the risk of recurrence between surgery and other first-line treatments overall, and stratified by sex and smoking status. These results are presented in Figure 4. As shown, the overall difference in risk of recurrence between first-line therapies was estimated to vary over time, with a 5.7% difference at one year, attenuating to 1.9% after five years. Stratified by patient

Table 3: Characteristics of the $n = 7,403$ patients in the Boston Lung Cancer Study cohort, overall and stratified by stage at diagnosis.

Characteristic	**Overall, $n = 7,403$**[1]	**Stage at Diagnosis**	
		1-3A, $n = 4,700$[1]	**3B-4**, $n = 2,703$[1]
First-Line Treatment			
Chemotherapy	1,851 (28%)	365 (8.0%)	1,486 (70%)
Other	7 (0.1%)	2 (<0.1%)	5 (0.2%)
Radiation	366 (5.5%)	194 (4.3%)	172 (8.1%)
Surgery	4,444 (67%)	3,994 (88%)	450 (21%)
Unknown	735	145	590
Age at Diagnosis (yrs.)	66 (59, 74)	68 (61, 74)	64 (56, 72)
Body Mass Index	26.4 (23.0, 31.1)	26.6 (23.3, 31.1)	25.7 (22.6, 30.1)
Sex			
Male	3,431 (46%)	2,093 (45%)	1,338 (50%)
Female	3,966 (54%)	2,603 (55%)	1,363 (50%)
Unknown	6 (<0.1%)	4 (<0.1%)	2 (<0.1%)
Race			
White/Caucasian	6,834 (92%)	4,349 (93%)	2,485 (92%)
Other	364 (4.9%)	212 (4.5%)	152 (5.6%)
Unknown	205 (2.8%)	139 (3.0%)	66 (2.4%)
Ethnicity			
Non-Hispanic	6,410 (87%)	3,990 (85%)	2,420 (90%)
Hispanic	87 (1.2%)	57 (1.2%)	30 (1.1%)
Unknown	906 (12%)	653 (14%)	253 (9.4%)
Education			
Some Grade School	438 (5.9%)	276 (5.9%)	162 (6.0%)
Some High School	976 (13%)	589 (13%)	387 (14%)
High School Graduate	1,451 (20%)	946 (20%)	505 (19%)
Vocational/Technical School	279 (3.8%)	156 (3.3%)	123 (4.6%)
Some College or Associate's Degree	1,469 (20%)	940 (20%)	529 (20%)
College Graduate	962 (13%)	604 (13%)	358 (13%)
Graduate or Professional School	831 (11%)	514 (11%)	317 (12%)
Other	997 (13%)	675 (14%)	322 (12%)
Smoking Status			
Never Smoker	1,009 (14%)	592 (13%)	417 (15%)
Former Smoker	4,251 (57%)	2,821 (60%)	1,430 (53%)
Current Smoker	1,979 (27%)	1,171 (25%)	808 (30%)
Smoker, Status Unknown	164 (2.2%)	116 (2.5%)	48 (1.8%)
Pack-Years of Smoking	40 (16, 53)	40 (19, 53)	37 (12, 54)
EGFR Mutation			
No	1,255 (17%)	737 (16%)	518 (19%)
Yes	298 (4.0%)	158 (3.4%)	140 (5.2%)
Not Tested	5,850 (79%)	3,805 (81%)	2,045 (76%)
KRAS Mutation			
No	1,148 (16%)	630 (13%)	518 (19%)
Yes	405 (5.5%)	265 (5.6%)	140 (5.2%)
Not Tested	5,850 (79%)	3,805 (81%)	2,045 (76%)

[1]n (%); Median (IQR)

sex, we see that among male patients, the risk difference is slightly higher, with a one-year difference of 5.9, attenuating to 2.0%, as compared to female patients, among whom we estimated the risk difference to be between 5.6% and 1.3% over five years. Larger differences were observed when stratifying by patient smoking status. As shown, treatment differences were slightly higher among current smokers, ranging from 5.9% to 2.5%, while among former (range: 5.6% to 1.2%) and never smokers (range: 5.6% to 0.1%) these differences were less.

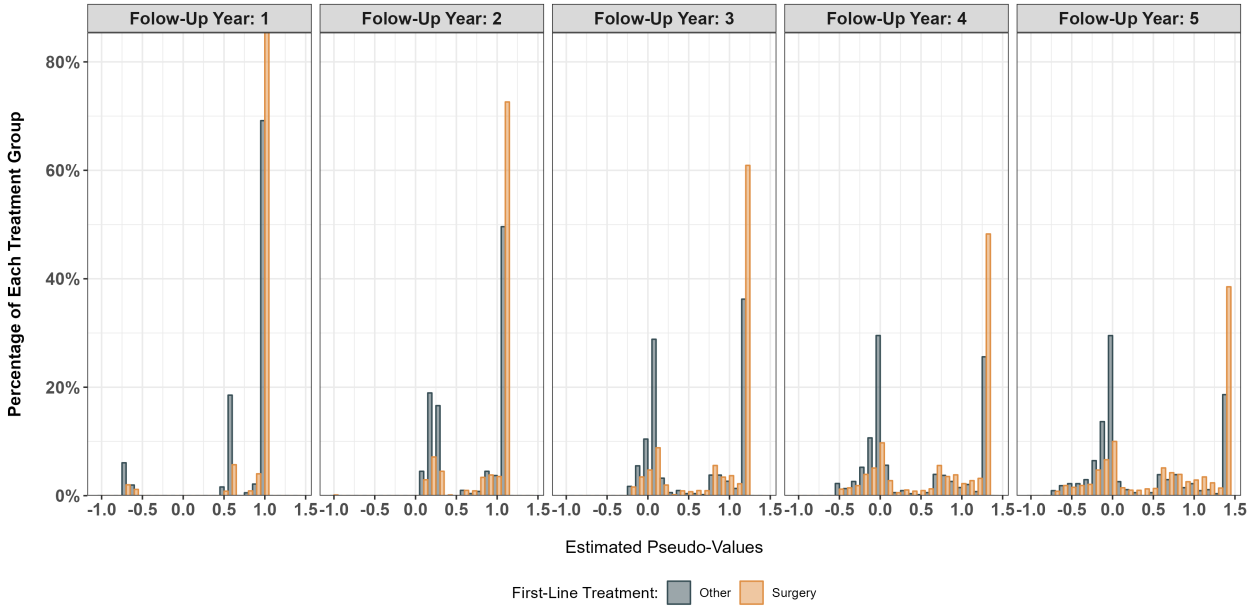
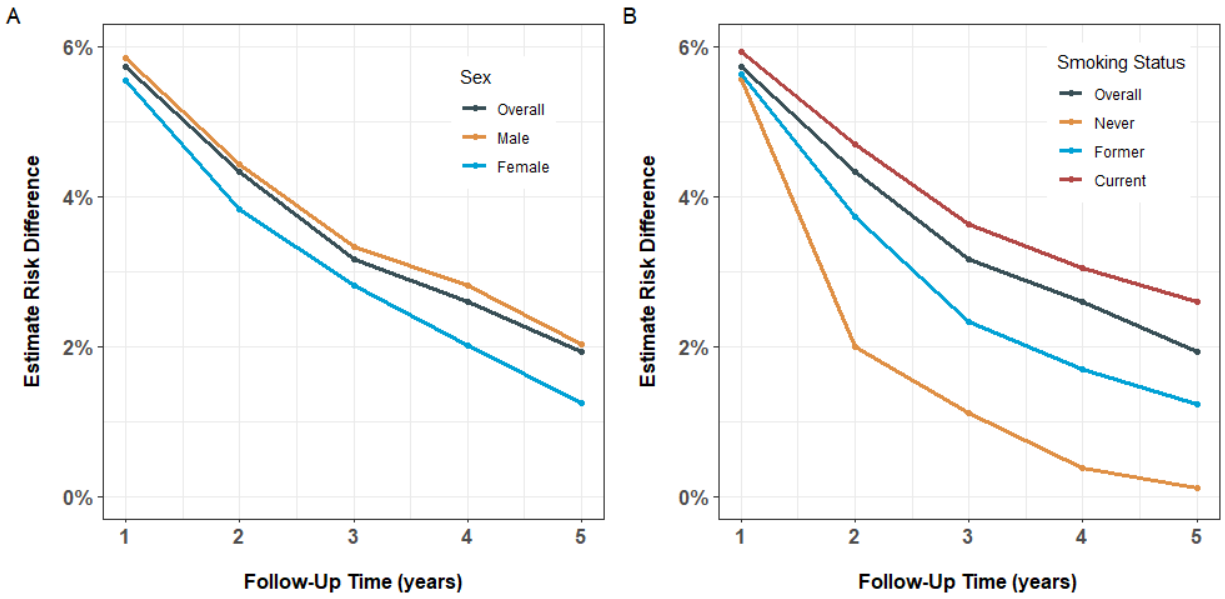


Figure 3: Estimated pseudo-survival probabilities for one-five years of follow-up, stratified by first-line treatment group.

Figure 4: Estimated average causal difference in the risk of recurrence between surgery and other first-line treatments among patients with stage 1-3A non-small cell lung cancer, over time and (A) stratified by sex; (B) stratified by smoking status



5 Discussion

In this work, we propose a deep learning framework for causal inference in time-to-event data with dependent censoring due to semi-competing risks, with a focus on non-fatal events such as time-to-recurrence. We demonstrate the performance of our approach on simulated data and apply it to a real-world dataset from a large epidemiologic lung cancer cohort. Our findings highlight the importance of accounting for semi-competing risks and provide new insights into the causal relationship between first line surgical resection and and the risk of recurrence. As shown, this approach provides an accurate method for estimating the causal average treatment effect on the probability of disease recurrence,

particularly in settings where the true relationship between the non-fatal outcome, treatment, and other confounding variables is complex.

Causal inference with time-to-event outcomes has received significant attention in the past decade. Traditional methods, such as the Cox Q-model, are commonly used in this context. However, these methods have several assumptions, which may not hold in complex real-world scenarios. For example, a common assumption is proportional hazards, or that the hazard ratio between two groups is constant over time. In practice, the hazard ratio may change over time due to time-varying confounding or effect modification. In addition, such approaches may not be able to handle high-dimensional data, where the number of covariates is much larger than the number of observations, or complicated functions of these potential risk factors. Machine learning methods have emerged as a promising alternative to handle the non-linear and non-proportional relationship between covariates and survival outcomes. Though not highlighted in this work, parametric and semi-parametric models requires extensive feature engineering and domain knowledge to select and transform predictors appropriately, whereas deep learning is scalable and flexible, can automatically learn relevant features from raw data, reducing the need for manual feature engineering, and is easily adaptable to various data types, including images and text. One application of machine learning to causal inference in survival analysis is targeted maximum likelihood estimation (TMLE) Stitelman et al. [2012], Zhu and Gallego [2020]. The TMLE approach uses a machine learning algorithm to estimate the propensity score, and then constructs a doubly robust estimator of the causal effect. Another popular approach is the use of random forests to estimate the survival function. Ishwaran et al. (2008) proposed the random survival forest algorithm, which is an extension of random forests for survival analysis, which was later extended by Cui et al. (2023) to causal survival forests, which estimate heterogeneous treatment effects with right-censored data Cui et al. [2023]. The random survival forest and causal survival forest algorithms have both been shown to outperform traditional methods in terms of estimation accuracy. In recent years, there has been a surge of interest in using deep learning methods for survival analysis [Katzman et al., 2018, Lee et al., 2018]. A promising aspect of deep learning is its ability to circumvent the curse of dimensionality in nonparametric settings by projecting the data into lower relevant representational space Bauer and Kohler [2019], Poggio et al. [2017], Abrol et al. [2021], Goodfellow et al. [2016]. For example, Katzman et al. (2018) proposed a deep learning framework called DeepSurv, which is a personalized treatment recommender system that uses a Cox proportional hazards deep neural network to predict the survival outcome of a patient given their clinical features via a multi-layer perceptron to estimate the hazard function. They showed that DeepSurv outperforms traditional survival models on several benchmark datasets. Another deep learning approach is the use of convolutional neural networks (CNNs) to extract features from the covariates. Lee et al. (2018) proposed a CNN-based survival analysis method called DeepHit, which learns a joint representation of the covariates and the time-to-event outcomes in competing risks settings.

From a different perspective, Zhao and Feng (2020) proposed the use of jackknife pseudo-values as targets for a deep neural network. The jackknife is a popular resampling technique used to estimate the influence of individual observations on a statistical estimator, which has been widely used in survival analysis to identify influential observations and assess model stability [Miller, 1974]. Recently, jackknife pseudo-values have been proposed for use as outcomes in survival regression settings [Andersen et al., 2003, 2004, Orenti et al., 2021], as well as target values for deep learning approaches [Zhao and Feng, 2020]. Recently, pseudo-outcomes have been proposed for causal survival analysis as well [Andersen et al., 2017]. We explored the applicability of pseudo-values to deep causal learners with dependent censoring. Causal inference in survival analysis with dependent censoring due to death is a challenging problem for several reasons. One of the main limitations of current approaches is the assumption of non-informative censoring, which assumes that censoring is independent of the survival outcome and the exposure of interest. This assumption does not hold in practice, as recurrence is often a strong precursor to death. Inappropriate handling of this dependence leads to biased estimates of causal effects.

A specific aim of this study was to focus on the effect of treatment on time to recurrence, rather than alternatives such as overall survival or progression-free survival, for several reasons. First, time to recurrence provides a more precise and clinically meaningful measure of the duration of response to treatment. Time to recurrence measures the time from diagnosis to the point where disease progression is observed, while composites such as progression-free survival measures the time to either disease progression or death. As a result, time to recurrence can more accurately capture the effect of treatment on disease progression, while progression-free survival can be confounded by the effect of treatment on survival. As remaining treatment options are dictated by the monitoring of disease progression, directly studying recurrence is less susceptible to bias than progression-free survival [Zappa and Mousa, 2016, Fedor et al., 2013].

In the context of the Boston Lung Cancer Study data, we observed differences in the efficacy of surgical resection compared to other first-line therapies, which attenuated over time. While there is limited literature on this topic, several studies suggest that surgical resection has better prognostic outcomes in patients with stage 1-3A NSCLC, particularly in the first five years of follow up [Wright et al., 2006, Uramoto and Tanaka, 2012]. Further, advances in surgical techniques have led to safer, less invasive procedures, which make surgery an important intervention, potentially in addition to other therapeutic regimens [Mitsudomi et al., 2013]. Additional, we note a modest difference in the effect of

surgery versus other first-line therapies when comparing male and female patient subgroups. While previous studies have reported similar rates of recurrence between these sub-populations [Keller et al., 2002], the timing of recurrence differs [Demicheli et al., 2012]. There is also evidence that female patients have a significantly better response to neoadjuvant chemotherapy than male patients [Cerfolio et al., 2006]. With respect to smoking status, we note many other individualized factors may contribute to greater perceived treatment benefits for current smokers versus former or never smokers, including stage and genetic mutations [Cortellini et al., 2021, Popat et al., 2022], warranting further study. Further, much of the literature on NSCLC prognosis points to a lack of emphasis on predictors of other clinical endpoints besides overall survival [Brundage et al., 2002]. Namely, research has shown that patients and providers are interested in endpoints such as disease recurrence and response to therapy, which impact quality of life and guide treatment decisions [Davidson et al., 1999].

We note some limitations of our proposed method and areas of future research. First, our workflow relies on estimating the marginal survival function for the non-fatal event time. In doing so, we utilize a plug-in estimator of our copula dependence parameter, θ, based on the extended concordance approach of Fine et al. (2001) [Fine et al., 2001, Orenti et al., 2021]. While this original, ad-hoc approach is shown to have good theoretical properties, in practice, correlations between covariates and higher-dimensional feature sets may incur biases. We offer a KNN-based approximation which addresses the impact of covariates on $\hat{\theta}$ while minimizing information leakage, however future work would employ more contemporary techniques for causal machine learning such as the recent 'cross-fitting' approach to causal estimation [Chernozhukov et al., 2018, 2022]. Second, the neural network architecture presented here is that of an S- or 'single' learner, in that the representational function of the feature space is learned for both treatment arms in the same architecture. The benefit of this approach suggested by previous work is that deep S-learners can transform covariates in a representation space that balances the covariate distribution between treated and control subjects [Johansson et al., 2016]. However, more sophisticated learners have been proposed which may further increase the utility of our proposed workflow. For example, T-learners utilize separate sub-network architectures to model each outcome separately, while treatment agnostic regression networks (TARNets) combine S- and T-learning by first encoding shared representation layers before training separate sub-networks for each treatment arm [Shalit et al., 2017, Koch et al., 2021]. Future work will explore these approaches and other algorithms such as targeted maximum likelihood estimation (TMLE) [Stitelman et al., 2012]. In addition, doubly robust methods may provide increased accuracy and efficiency as they require only one of the outcome model or a treatment model to be correctly estimated [Hu et al., 2021a, Steingrimsson and Morrison, 2020]. Lastly, previous work has shown that, in the context of time-to-event endpoints, using imputed outcomes in a deep learning framework is asymptotically more efficient than directly optimizing a loss function of the observed survival times [Steingrimsson and Morrison, 2020]. Further study of these results in the context of our approach may provide additional justification for our method.

There are also several open problems and areas of future direction. A primary concern is how to conduct inference in this setting. Our approach yields accurate point estimates for our causal estimand, but we do not yet have a means of quantifying the uncertainty surrounding these estimates. While uncertainty quantification in causal deep learning is still relatively new, it is an important step in developing methods that have practical clinical applicability [Abdar et al., 2021]. Other approaches such as Bayesian neural networks may lead to valid inference for testing for the significance of the causal effect estimates. Further, the implementation is computationally intensive, owing to the intermediate steps needed to calculate the marginal survival functions and pseudo-responses before training our deep neural network. Future work will improve the efficiency of the proposed method. We also consider extending this approach to other useful target values, such as restricted mean survival times, and to other diseases, such as renal disease [Feng et al., 2019], which may also yield similar data structures. We will address these problems in subsequent work. Overall, however, we demonstrate the performance of this approach on simulated and real-world data, highlighting its ability to accurately estimate the causal effect in the presence of semi-competing risks. Our findings demonstrate the importance of accounting for dependent censoring due to semi-competing risks when estimating the causal effect of treatment on time-to-non fatal events.

Disclosure Statement

The authors declare no conflicts of interest.

Funding

National Institutes of Health grant R01CA249096 (YL)

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

Anees Abrol, Zening Fu, Mustafa Salman, Rogers Silva, Yuhui Du, Sergey Plis, and Vince Calhoun. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature communications*, 12(1):1–17, 2021.

Kwang Woo Ahn and Franco Mendolia. Pseudo-value approach for comparing survival medians for dependent data. *Statistics in medicine*, 33(9):1531–1538, 2014.

JJ Allaire and François Chollet. *keras: R Interface to 'Keras'*, 2022. URL `https://CRAN.R-project.org/package=keras`. R package version 2.9.0.

JJ Allaire and Yuan Tang. *tensorflow: R Interface to 'TensorFlow'*, 2022. URL `https://CRAN.R-project.org/package=tensorflow`. R package version 2.9.0.

Per K Andersen, Elisavet Syriopoulou, and Erik T Parner. Causal inference in survival analysis using pseudo-observations. *Statistics in medicine*, 36(17):2669–2681, 2017.

Per Kragh Andersen and Maja Pohar Perme. Pseudo-observations in survival analysis. *Statistical methods in medical research*, 19(1):71–99, 2010.

Per Kragh Andersen, John P Klein, and Susanne Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, 2003.

Per Kragh Andersen, Mette Gerster Hansen, and John P Klein. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime data analysis*, 10(4):335–350, 2004.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

Michael D Brundage, Diane Davies, and William J Mackillop. Prognostic factors in non-small cell lung cancer: a decade of progress. *Chest*, 122(3):1037–1057, 2002.

Robert James Cerfolio, Ayesha S Bryant, Ethan Scott, Manisha Sharma, Francisco Robert, Sharon A Spencer, and Robert I Garver. Women with pathologic stage i, ii, and iii non-small cell lung cancer have better survival than men. *Chest*, 130(6):1796–1802, 2006.

Pei-Yun Chen and Anastasios A Tsiatis. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57(4):1030–1038, 2001.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022.

David C Christiani. The boston lung cancer survival cohort, 2017.

David G Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

Leah Comment, Fabrizia Mealli, Sebastien Haneuse, and Corwin Zigler. Survivor average causal effects for continuous time: a principal stratification approach to causal inference with semicompeting risks. *arXiv preprint arXiv:1902.09304*, 2019.

Alessio Cortellini, Andrea De Giglio, Katia Cannita, Diego L Cortinovis, Robin Cornelissen, Cinzia Baldessari, Raffaele Giusti, Ettore D'Argento, Francesco Grossi, Matteo Santoni, et al. Smoking status during first-line immunotherapy and chemotherapy in nsclc patients: A case–control matched analysis from a large multicenter study. *Thoracic Cancer*, 12(6):880–889, 2021.

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

Lei Cui, Hansheng Li, Wenli Hui, Sitong Chen, Lin Yang, Yuxin Kang, Qirong Bo, and Jun Feng. A deep learning-based framework for lung cancer survival analysis with biomarker interpretation. *BMC bioinformatics*, 21:1–14, 2020.

Yifan Cui, Michael R Kosorok, Erik Sverdrup, Stefan Wager, and Ruoqing Zhu. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):179–211, 2023.

Judith R Davidson, Michael D Brundage, and Deb Feldman-Stewart. Lung cancer treatment decisions: patients' desires for participation and information. *Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer*, 8(6):511–520, 1999.

Amanda Delgado and Achuta Kumar Guddati. Clinical endpoints in oncology-a primer. *American journal of cancer research*, 11(4):1121, 2021.

Romano Demicheli, Marco Fornili, Federico Ambrogi, Kristin Higgins, Jessamy A Boyd, Elia Biganzoli, and Chris R Kelsey. Recurrence dynamics for non–small-cell lung cancer: effect of surgery on the development of metastases. *Journal of Thoracic Oncology*, 7(4):723–730, 2012.

Shreyesh Doppalapudi, Robin G Qiu, and Youakim Badr. Lung cancer survival period prediction and understanding: Deep learning approaches. *International Journal of Medical Informatics*, 148:104371, 2021.

David Fedor, W Rainey Johnson, and Sunil Singhal. Local recurrence following lung cancer surgery: incidence, risk factors, and outcomes. *Surgical oncology*, 22(3):156–161, 2013.

Yanhuan Feng, Rongshuang Huang, Janet Kavanagh, Lingzhi Li, Xiaoxi Zeng, Yi Li, and Ping Fu. Efficacy and safety of dual blockade of the renin–angiotensin–aldosterone system in diabetic kidney disease: A meta-analysis. *American Journal of Cardiovascular Drugs*, 19:259–286, 2019.

Jason P Fine, Hongyu Jiang, and Rick Chappell. On semi-competing risks data. *Biometrika*, 88(4):907–919, 2001.

Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

Debashis Ghosh. A causal framework for surrogate endpoints with semi-competing risks data. *Statistics & probability letters*, 82(11):1898–1902, 2012.

Emilio AL Gianicolo, Martin Eichler, Oliver Muensterer, Konstantin Strauch, and Maria Blettner. Methods for evaluating causality in observational studies: Part 27 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, 117(7):101, 2020.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Miguel A Hernán and James M Robins. Causal inference, 2010.

Jin-Jian Hsieh and Yu-Ting Huang. Regression analysis based on conditional likelihood approach under semi-competing risks data. *Lifetime data analysis*, 18(3), 2012.

Liangyuan Hu, Jiayi Ji, and Fan Li. Estimating heterogeneous survival treatment effect in observational data using machine learning. *Statistics in medicine*, 40(21):4691–4713, 2021a.

Liangyuan Hu, Jung-Yi Lin, Keith Sigel, and Minal Kale. Estimating heterogeneous survival treatment effects of lung cancer screening approaches: A causal machine learning analysis. *Annals of Epidemiology*, 62:36–42, 2021b.

Yen-Tsung Huang. Causal mediation of semicompeting risks. *Biometrics*, 77(4):1143–1154, 2021.

Ina Jazić, Deborah Schrag, Daniel J Sargent, and Sebastien Haneuse. Beyond composite endpoints analysis: semicompeting risks as an underutilized framework for cancer research. *JNCI: Journal of the National Cancer Institute*, 108 (12), 2016.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245): 255–260, 2015.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deep survival: A deep cox proportional hazards network. *stat*, 1050(2):1–10, 2016.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):1–12, 2018.

SM Keller, MG Vangel, S Adak, H Wagner, JH Schiller, A Herskovic, R Komaki, MC Perry, RS Marks, RB Livingston, et al. The influence of gender on survival and tumor recurrence following adjuvant therapy of completely resected stages ii and iiia non-small cell lung cancer. *Lung cancer*, 37(3):303–309, 2002.

Ruth H Keogh, Jon Michael Gran, Shaun R Seaman, Gwyneth Davies, and Stijn Vansteelandt. Causal inference in survival analysis using longitudinal observational data: Sequential trials and marginal structural models. *arXiv preprint arXiv:2110.03117*, 2021.

Sehee Kim, Donglin Zeng, Lloyd Chambless, and Yi Li. Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in biosciences*, 4:262–281, 2012.

Bernard Koch, Tim Sainburg, Pablo Geraldo, Song Jiang, Yizhou Sun, and Jacob Gates Foster. Deep learning of potential outcomes. *arXiv preprint arXiv:2110.04442*, 2021.

Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Wenhua Liang, Jun Liu, and Jianxing He. Driving the improvement of lung cancer prognosis. *Cancer Cell*, 38(4): 449–451, 2020.

Brent R Logan, Mei-Jie Zhang, and John P Klein. Marginal models for clustered time-to-event data with competing risks using pseudovalues. *Biometrics*, 67(1):1–7, 2011.

Dustin M Long and Michael G Hudgens. Sharpening bounds on principal effects with covariates. *Biometrics*, 69(4): 812–819, 2013.

Rupert G Miller. The jackknife-a review. *Biometrika*, 61(1):1–15, 1974.

Tetsuya Mitsudomi, Kenichi Suda, and Yasushi Yatabe. Surgery for nsclc in the era of personalized medicine. *Nature reviews Clinical oncology*, 10(4):235–244, 2013.

Daniel Nevo and Malka Gorfine. Causal inference for semi-competing risks data. *Biostatistics*, 23(4):1115–1132, 2022.

David Oakes. Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84(406): 487–493, 1989.

Annalisa Orenti, Patrizia Boracchi, Giuseppe Marano, Elia Biganzoli, and Federico Ambrogi. A pseudo-values regression model for non-fatal event free survival in the presence of semi-competing risks. *Statistical Methods & Applications*, pages 1–19, 2021.

Judea Pearl. Causal inference in statistics: An overview. *Statistical Surveys*, 3:96–146, 2009.

Judea Pearl. Principal stratification–a goal or a tool? *The international journal of biostatistics*, 7(1):1–13, 2011.

Limin Peng and Jason P Fine. Regression modeling of semicompeting risks data. *Biometrics*, 63(1):96–108, 2007.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

Katerina Politi and Roy S Herbst. Lung cancer in the era of precision medicine. *Clinical cancer research*, 21(10): 2213–2220, 2015.

Sanjay Popat, Stephen V. Liu, Nicolas Scheuer, Alind Gupta, Grace G. Hsu, Sreeram V. Ramagopalan, Frank Griesinger, and Vivek Subbiah. Association Between Smoking History and Overall Survival in Patients Receiving Pembrolizumab for First-Line Treatment of Advanced Non–Small Cell Lung Cancer. *JAMA Network Open*, 5(5):e2214046–e2214046, 05 2022.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. In *Machine Learning for Healthcare Conference*, pages 101–114. PMLR, 2016.

Denise Rava. *Survival Analysis and Causal Inference: from Marginal Structural Cox to Additive Hazards Model and beyond*. University of California, San Diego, 2021.

T Richardson and A Rotnitzky. Causal etiology of the research of james m. robins. *Statistical Science*, 29(4):459–484, 2014.

Federico Rotolo, Catherine Legrand, and Ingrid Van Keilegom. A simulation procedure based on copulas to generate clustered multi-state survival data. *Computer methods and programs in biomedicine*, 109(3):305–312, 2013.

David Ruppert and David S Matteson. *Statistics and data analysis for financial engineering*, volume 13. Springer, 2011.

Camille Sabathé, Per K Andersen, Catherine Helmer, Thomas A Gerds, Hélène Jacqmin-Gadda, and Pierre Joly. Regression analysis in an illness-death model with interval-censored data: A pseudo-value approach. *Statistical methods in medical research*, 29(3):752–764, 2020.

Stephen Salerno and Yi Li. Deep learning of semi-competing risk data via a new neural expectation-maximization algorithm, 2022.

Stephen Salerno and Yi Li. High-dimensional survival analysis: Methods and applications. *arXiv*, Preprint posted online May 5, 2022. arXiv:2205.02948 [stat.ME]. doi: 10.48550/arXiv.2205.02948.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, and Ahmedin Jemal. Cancer statistics, 2023. *CA: a cancer journal for clinicians*, 73(1):17–48, 2023.

Raphael Sonabend, Franz J Király, Andreas Bender, Bernd Bischl, and Michel Lang. mlr3proba: an r package for machine learning in survival analysis. *Bioinformatics*, 37(17):2789–2791, 2021.

Jon Arni Steingrimsson and Samantha Morrison. Deep learning for survival outcomes. *Statistics in medicine*, 39(17): 2339–2349, 2020.

Matthew A Steliga and Carolyn M Dresler. Epidemiology of lung cancer: smoking, secondhand smoke, and genetics. *Surgical Oncology Clinics*, 20(4):605–618, 2011.

Ori M Stitelman, Victor De Gruttola, and Mark J van der Laan. A general implementation of tmle for longitudinal data applied to causal inference in survival analysis. *The international journal of biostatistics*, 8(1), 2012.

Eric J Tchetgen Tchetgen. Identification and estimation of survivor average causal effects. *Statistics in medicine*, 33 (21):3601–3628, 2014.

Hidetaka Uramoto and Fumihiro Tanaka. Prediction of recurrence after complete resection in patients with nsclc. *Anticancer research*, 32(9):3953–3960, 2012.

Ashley J Vargas and Curtis C Harris. Biomarker development in the precision medicine era: lung cancer as a case study. *Nature Reviews Cancer*, 16(8):525–537, 2016.

Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.

Gavin Wright, Renee L Manser, Graham Byrnes, David Hart, and Donald A Campbell. Surgery for non-small cell lung cancer: systematic review and meta-analysis of randomised controlled trials. *Thorax*, 61(7):597–603, 2006.

Yujiao Wu, Jie Ma, Xiaoshui Huang, Sai Ho Ling, and Steven Weidong Su. Deepmmsa: A novel multimodal deep learning method for non-small cell lung cancer survival analysis. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1468–1472. IEEE, 2021.

Jinfeng Xu, John D Kalbfleisch, and Beechoo Tai. Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics*, 66(3):716–725, 2010.

Yanxun Xu, Daniel Scharfstein, Peter Müller, and Michael Daniels. A bayesian nonparametric approach for evaluating the causal effect of treatment in randomized trials with semi-competing risks. *Biostatistics*, 23(1):34–49, 2022.

Jiawen Yao, Xinliang Zhu, Feiyun Zhu, and Junzhou Huang. Deep correlational learning for survival prediction from multi-modality data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 406–414. Springer, 2017.

Cecilia Zappa and Shaker A Mousa. Non-small cell lung cancer: current treatment and future advances. *Translational lung cancer research*, 5(3):288, 2016.

Tamir Zehavi and Daniel Nevo. Matching methods for truncation by death problems. *arXiv preprint arXiv:2110.10186*, 2021.

Junni L Zhang and Donald B Rubin. Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4):353–368, 2003.

Lili Zhao and Dai Feng. Deep neural networks for survival analysis using pseudo values. *IEEE Journal of Biomedical and Health Informatics*, 24(11):3308–3314, 2020.

Qingyuan Zhao and Trevor Hastie. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281, 2021.

Jie Zhu and Blanca Gallego. Targeted estimation of heterogeneous treatment effect in observational survival analysis. *Journal of Biomedical Informatics*, 107:103474, 2020.

Blaž Zupan, Janez Demšar, Michael W Kattan, J Robert Beck, and Ivan Bratko. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial intelligence in medicine*, 20(1):59–75, 2000.

A Supplemental Simulation Information

A.1 Additional Simulation Details

Table A1 below lists the settings for our main simulation and sensitivity analysis.

Table A1: Simulation Settings

Setting	Sample Size	θ	τ	Censoring Rate	Data Generating Mechanism
Main Simulations					
1	500	0.5	0.2	0%	Proportional Hazards, Linear Risk
2	1,000	0.5	0.2	0%	Proportional Hazards, Linear Risk
3	500	2.0	0.5	0%	Proportional Hazards, Linear Risk
4	1,000	2.0	0.5	0%	Proportional Hazards, Linear Risk
5	500	0.5	0.2	50%	Proportional Hazards, Linear Risk
6	1,000	0.5	0.2	50%	Proportional Hazards, Linear Risk
7	500	2.0	0.5	50%	Proportional Hazards, Linear Risk
8	1,000	2.0	0.5	50%	Proportional Hazards, Linear Risk
9	500	0.5	0.2	0%	Proportional Hazards, Non-Linear Risk
10	1,000	0.5	0.2	0%	Proportional Hazards, Non-Linear Risk
11	500	2.0	0.5	0%	Proportional Hazards, Non-Linear Risk
12	1,000	2.0	0.5	0%	Proportional Hazards, Non-Linear Risk
13	500	0.5	0.2	50%	Proportional Hazards, Non-Linear Risk
14	1,000	0.5	0.2	50%	Proportional Hazards, Non-Linear Risk
15	500	2.0	0.5	50%	Proportional Hazards, Non-Linear Risk
16	1,000	2.0	0.5	50%	Proportional Hazards, Non-Linear Risk
Sensitivity Analysis					
17	500	2.0	-	0%	Proportional Hazards, Linear Risk, Gumbel Copula
18	1,000	2.0	-	0%	Proportional Hazards, Linear Risk, Gumbel Copula
19	500	2.0	-	50%	Proportional Hazards, Linear Risk, Gumbel Copula
20	1,000	2.0	-	50%	Proportional Hazards, Linear Risk, Gumbel Copula

A.2 Data Generation Procedures

In the following, we detail the the data generation procedure for our simulation studies.

Proportional Hazards Model, Linear Risk Function

Given the formulation of the Clayton copula, we can express the bivariate surival function of the non-fatal, T_{i1}, and fatal, T_{i2}, event times as

$$S(t_1, t_2) = \Pr(T_{i1} > t_1, T_{i2} > t_2) = \left[S_1(t_1)^{-\theta} + S_2(t_2)^{-\theta} - 1\right]^{-\frac{1}{\theta}}; 0 \leq t_1 \leq t_2,$$

where $S_1(t_1)$ is the marginal survival function of the non-fatal event, $S_2(t_2)$ is the marginal survival function of the fatal event, and θ is the copula parameter which measures the dependence between the non-fatal and fatal event times. In the first simulation, we generated non-fatal (T_{i1}) and fatal (T_{i2}) event times from marginal models specified by

$$\log(T_{i1}/3) = -(\beta_1 Z_i + \beta_1 X_{i1} + \beta_1 X_{i2}) + \varepsilon_{i1}$$
$$\log(T_{i2}/3) = -(\beta_2 Z_i + \beta_2 X_{i1} + \beta_2 X_{i2}) + \varepsilon_{i2},$$

where Z_i is a Bernoulli random variable with a success probability of 0.5, X_{i1} and X_{i12} are independent truncated normal random variables with mean 1, variance 0.5, and truncation bounds of [0, 2], and $(\varepsilon_{i1}, \varepsilon_{i2})$ are correlated random errors. To induce dependence between the event times, we simulate ε_{i1} and ε_{i2} from the Clayton copula model,

$$\left[\Pr\left(\varepsilon_{i1} > t_1\right)^{-\theta} + \Pr\left(\varepsilon_{i2} > t_2\right)^{-\theta} - 1\right]^{-\frac{1}{\theta}},$$

where ε_{i1} and ε_{i2} follow the extreme value distribution, i.e., $\Pr(\varepsilon_{i1} > t_1) = \exp\{-\exp(t_1)\}$ and $\Pr(\varepsilon_{i2} > t_2) = \exp\{-\exp(t_2)\}$ [see Rotolo et al., 2013]. The data generation procedure is as follows:

1. Draw two independent uniform random variables, $U_{i1}, V_{i2} \sim \text{Unif}(0,1)$
2. Set $\varepsilon_{i1} = \log\{-\log(U_{i1})\}$
3. Set $U_{i2} = \left[\left(V_{i2}^{-\theta/(1+\theta)} - 1\right) \times \exp\{\theta \exp(\varepsilon_{i1})\} + 1\right]^{-1/\theta}$
4. Set $\varepsilon_{i2} = \log\{-\log(U_{i2})\}$
5. Draw a Bernoulli random variable, Z_i, with success probability 0.5
6. Draw X_{i1}, X_{i2} from independent N(1, 0.5) distributions with truncation bounds [0, 2]
7. Set $T_{i1} = 3 \times \exp\{-(\beta_1 Z_i + \beta_1 X_{i1} + \beta_1 X_{i2}) + \varepsilon_{i1}\}$ with $\beta_1 = 1$
8. Set $T_{i2} = 3 \times \exp\{-(\beta_2 Z_i + \beta_2 X_{i1} + \beta_2 X_{i2}) + \varepsilon_{i2}\}$ with $\beta_2 = 0.2$
9. Draw C_i, from a mixture of uniforms, where $C_i \sim \xi_i \text{Unif}(0,1) + (1-\xi_i)\text{Unif}(1, 1.2)$ with $\xi_i \sim \text{Bern}(0.2)$
10. Set $Y_{i2} = \min(T_{i2}, C_i)$, $\delta_{i2} = \mathbb{I}(T_{i2} \leq C_i)$, $Y_{i1} = \min(T_{i1}, Y_{i2})$, $\delta_{i1} = \mathbb{I}(T_{i1} \leq Y_{i2})$
11. Repeat steps (1) - (10) for $i = 1, \ldots, n$
12. Return $\{(Y_{i1}, \delta_{i1}, Y_{i2}, \delta_{i2}, Z_i, X_{i1}, X_{i2}); i = 1, \ldots, n\}$

Proportional Hazards Model, Non-Linear Risk Function

In this setting, we repeat the same data generation procedure as listed above, except

- In step (6), we draw X_{i1}, X_{i2} from independent N(0, 0.5) distributions with truncation bounds [-1, 1]
- In step (7), we set $T_{i1} = 3 \times \exp\{-(\beta_1 Z_i + \beta_1 X_{i1}^2 + \beta_1 X_{i2}^2) + \varepsilon_{i1}\}$
- In step (8), we set $T_{i2} = 3 \times \exp\{-(\beta_2 Z_i + \beta_2 X_{i1}^2 + \beta_2 X_{i2}^2) + \varepsilon_{i1}\}$.

Sensitivity Analysis: Gumbel Copula

In this setting, we repeat the same data generation procedure as the first simulation setting, except in step (3), we set

$$U_{i2} = \exp\left\{-\left[\{-\log(V_{i2})\}^{\theta} + [-\log(U_{i1})]^{\theta}\right]^{\frac{1}{\theta}}\right\}$$

A.3 Sensitivity Analysis Results

Table A2 below gives the results of the sensitivity analysis where we compare the performance of our proposed method under the assumed Clayton copula versus an alternative Gumbel copula. As shown, we do incur bias if our model is misspecified for the data generating copula, with the ATE tending to be underestimated for the Gumbel copula. We further see that we have a higher mean squared error across all settings when the data are generated from the Gumbel copula, as compared to the Clayton copula.

Table A2: Average bias and mean squared error (MSE) for estimated vs. true ATE comparing our proposed method under the Clayton versus Gumbel copula. Results are averaged over 50 independent datasets for each setting.

Simulation Settings			**Bias**		**Mean Squared Error**	
n	θ	**Censoring**	**Clayton**	**Gumbel**	**Clayton**	**Gumbel**
500	2.0	50%	0.0057	-0.0837	0.0053	0.0076
500	2.0	0%	0.0069	-0.0771	0.0011	0.0068
1000	2.0	50%	0.0048	-0.0756	0.0037	0.0065
1000	2.0	0%	0.0030	-0.0750	0.0021	0.0062