Aligning Machine Learning Solutions with Clinical Needs

by

Fahad Kamran

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2023

Doctoral Committee:

Associate Professor Jenna Wiens, Chair
Assistant Professor David Fouhey
Assistant Professor Rahul Ladhania
Assistant Professor Maggie Makar

Fahad Kamran
fhdkmrn@umich.edu
ORCID iD: 0000-0003-2488-8887

ACKNOWLEDGEMENTS

First, all praise belongs to Allah, The Most Gracious, The Most Merciful.

I would like to thank my committee members for their valuable feedback on this dissertation and throughout my journey toward the PhD. Specifically, I would like to thank my advisor Professor Jenna Wiens. I have been fortunate to spend over five years learning from Jenna and becoming a better researcher and person from it. She encouraged me through the hardest times and always supported my interests and endeavors, both inside of research and out. Without her support and guidance, my growth throughout my PhD would not have been possible.

Next, I would like to thank all of my collaborators and colleagues who have made my research journey so unique and special. In particular, I would like to thank all those in Mechanical Engineering and Michigan Medicine whose interdisciplinary collaborations have helped shape my approach to research and my interest in the field. Moreover, I would like to thank all of my wonderful colleagues at Evidation Health, especially Dr. Eric J. Daza who was an incredible mentor during my internship. Eric's advice has been instrumental in my professional and academic growth.

Next, I would like to thank all of the friends who have helped keep me afloat throughout my PhD. First, I would like to thank everyone in the MLD3 lab. I never expected that some of my best friends would be the people who worked in the lab with me. Second, I would like to thank all of the friends I made throughout my PhD who helped me in the good times and the bad times. I'm sure I'll miss some really important people on this list, but a special thank you to Kevin, Eli, Caleb, Sarah, Won, Trevor, and Mohamed.

Finally, I could not have finished my dissertation without the support and love of all of my family. First, thank you to Hassan and Saniya, my best friends who've helped me mature into the person I am today. Second, thank you to my wife Hina for all your support throughout my PhD and for keeping me sane through all of the most difficult times. Finally, thank you to Mama and Papa for your guidance throughout my whole life. I am forever grateful for all that you have done for me.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

LIST OF ACRONYMS

ML Machine Learning

AUROC Area Under the Receiver Operating Characteristic Curve

CATE Conditional Average Treatment Effect

DR Doubly Robust

PEHE Precision in Estimating Heterogeneous Treatment Effects

AUTOC Area Under the Targeting Operator Characteristic

DDC Distributional Divergence for Calibration

ESM Epic Sepsis Model

ABSTRACT

The availability of large observational datasets in healthcare presents an opportunity to leverage machine learning techniques to learn complex relationships between an individual's characteristics, underlying health status, and response to interventions. Despite progress, there is often a mismatch between how machine learning models are developed and clinical needs. In this dissertation, we study how considering clinical needs can and should inform model development in healthcare.

First, in survival analysis, deep learning approaches have been proposed for estimating an individual's survival probability over some time horizon. However, these methods often focus on optimizing discriminative performance and have ignored model calibration. Well-calibrated survival curves present realistic and meaningful probabilistic estimates of the true underlying survival process for an individual, an essential characteristic for survival analysis models in many clinical contexts. In light of the shortcomings of existing approaches, we propose a new training scheme for optimizing deep survival analysis models for strong discriminative performance and good calibration. Across two clinical datasets, we show that our approach yields models with strong discriminative performance while improving calibration over existing methods.

Second, in causal inference, past work has focused on accurately estimating conditional average treatment effects (CATEs) to help guide treatment allocation. However, in many settings, decision-makers only require a ranking of individuals to assist in allocating treatments. Leveraging the insight that ranking can be simpler than CATE estimation and better CATE accuracy doesn't necessarily translate to better treatment allocation, we propose an approach that optimizes directly for rankings of individuals to maximize benefit of treatment. Our tree-based approach maximizes the expected benefit across all treatment thresholds using a novel splitting criteria. Through experiments on synthetic datasets, we show that the proposed approach leads to better sample efficiency and better treatment assignments, as measured by expected benefit, compared to models optimized for accurate CATEs.

Third, when exact CATEs are needed, we study the mismatch between theoretical results in CATE estimation and how this theory holds empirically. In recent years, techniques incorporating estimates of both the propensity score and potential outcomes have gained

popularity in part due to their strong theoretical guarantees for overcoming confounding bias. However, how this theory translates to practice across an extensive set of practical settings, especially in the context of deep learning, has not been well explored. We present an in-depth exploration of popular techniques, finding that those relying only on estimates of the outcome, in particular the X-Learner, can consistently outperform more sophisticated techniques across a variety of practical settings.

Finally, we study how the mismatch between machine learning objectives and clinical needs manifests in existing clinical tools for sepsis risk stratification. Standard risk-stratification approaches focus on predicting the likelihood of sepsis before the sepsis criteria is met. However, both the training and evaluation of these models do not match the ultimate goal of augmenting clinical decision-making to improve patient outcomes. We study both challenges, finding that: 1) existing risk stratification approaches deteriorate significantly when evaluating before clinical recognition of sepsis and 2) targeting those most likely to develop sepsis may be sub-optimal with respect to improving patient outcomes.

Overall, our contributions bridge, in part, the gap between machine learning research and practice in healthcare. Ultimately, by recognizing domain-specific needs in clinical care as we have, machine learning practitioners can develop more impactful models.

CHAPTER 1

Introduction

In recent years, there has been tremendous growth in the availability of observational clinical data [167, 147, 205]. This growth has led to an exponential increase in research at the intersection of machine learning (ML) and healthcare [77, 62, 86, 50, 6, 154, 181, 206]. A primary goal in developing ML algorithms for healthcare is to augment clinicians' understanding of patient risk and improve patient care. ML models can achieve this goal in many ways, including assisting in risk-stratification and treatment allocation. Accordingly, two major subfields of research are focused on these goals. First, survival analysis is interested in understanding a patient's expected risk of an adverse event over time [133]. Second, causal effect estimation is interested in measuring the causal effect of interventions in patient care on patient risk [150, 78]. Together, work in these fields can help improve clinical decision support, towards providing timely interventions to those most in need.

The potential for ML to augment clinical care is exciting, but learning from clinical data presents many technical challenges. For example, censored individuals, for whom the presence of a potentially adverse event is unknown, make it difficult for models to estimate the risk of acquiring the event. Moreover, the presence of confounding in observational data makes it difficult to learn the causal effect of a treatment or intervention on a patient's outcome. To overcome these challenges, researchers in these fields have developed novel methods for learning ML models from observational data [214, 202].

Despite this progress, ML has had only a limited impact in clinical practice in these fields. We hypothesize that the lack of adoption is in part due to a mismatch between ML research and what is needed in certain clinical contexts (**Figure 1.1**). New methodologies in these fields are often optimized for and tested via metrics that may not represent how they may be used in clinical care. Hence, if and how these methods may augment clinical care are often not incorporated directly into the training and evaluation process. **Our central thesis is that specific clinical needs can and should inform training and evaluation of ML models for greater impact in clinical care.**

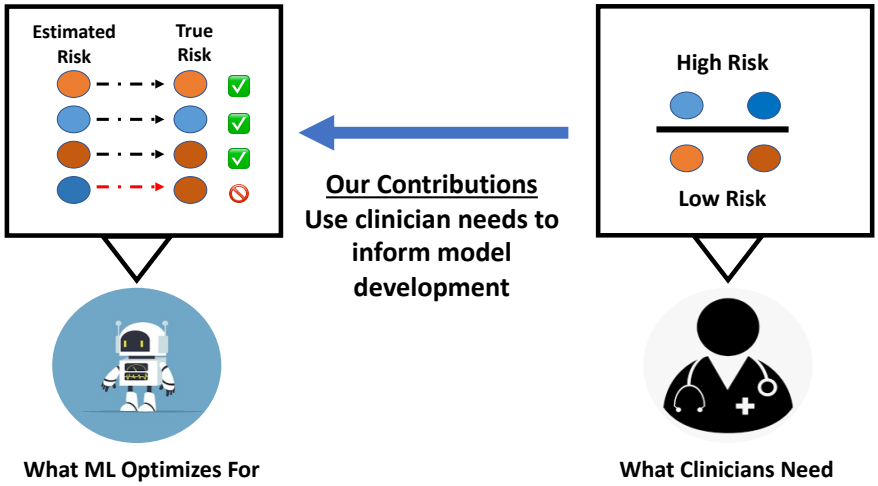

Figure 1.1: In this dissertation, we study the mismatch between what ML models are optimized and evaluated for and what the needs of healthcare are. Optimizing for a particular task, such as predicting estimated risk exactly, may not be aligned with how clinicians would want to use an ML model, such as for identifying low-risk and high-risk patients. We focus on aligning the objectives of ML research with the needs of healthcare. We consider how we can use clinician needs to inform model development to maximize impact and improve clinical care.

1.1 Challenges and Opportunities

In this dissertation, we develop and evaluate new ML techniques to bridge the disconnect between ML research and clinical needs for a greater impact in clinical care. We focus on the fields of survival analysis and causal effect estimation at an individual-level, with an ultimate eye towards precision medicine [73, 215, 11]. From here, we study how the gap between ML research and clinical needs manifests in the existing clinical tools for sepsis risk stratification. Below, we summarize each direction.

First, we consider the disconnect between progress in the field of survival analysis and the needs in clinical care. Recent work in survival analysis has focused on the use of deep learning approaches for estimating an individual's probability of survival over some time horizon [116, 156, 104]. Such approaches can capture complex non-linear relationships, without relying on restrictive assumptions regarding the relationship between an individual's characteristics and their underlying survival process. Moreover, the flexible nature of survival analysis allows custom loss functions that may affect different aspects of a learned survival model. To date, deep survival methods have focused primarily on optimizing discriminative performance and have ignored model calibration. Well-calibrated survival curves present realistic and

meaningful probabilistic estimates of the true underlying survival process for an individual, which may be used to better understand how an individual should be treated. However, due to the lack of ground-truth regarding the underlying stochastic process of survival for an individual, optimizing and measuring calibration in survival analysis is an inherently difficult task. In contrast to past work, in **Chapter 3**, we recognize the importance of calibration in clinical care and consider optimizing for and evaluating with respect to it when building survival analysis techniques. In particular, we: i) highlight the shortcomings of existing approaches in terms of calibration and ii) propose a new training scheme for optimizing deep survival analysis models that maximizes discriminative performance, subject to good calibration. We consider both theoretical and empirical justification to highlight our proposed methods for training and evaluating for calibration in survival analysis.

Next, we consider the field of estimating conditional average treatment effects (CATEs), or the causal effect of a treatment for an individual given their particular characteristics or covariates, from observational data [182, 90]. The ability to accurately estimate CATEs can help guide clinical decision-making and treatment allocation [57]. Assigning treatment based on a ranking of who is most likely to benefit from a particular resource or intervention, i.e., who has a higher estimated CATE, is a potential solution to the problem of resource allocation when the goal is to maximize overall benefit [27, 111, 140, 39]. Hence, current causal inference approaches for estimating CATEs often prioritize accuracy. However, in resource constrained-settings, decision makers may only need an accurate ranking of individuals to allocate treatments. In these scenarios, exact CATE estimation may be an unnecessarily challenging task, particularly when the underlying function is difficult to learn. Inaccurate or biased estimates can still lead to the optimal ordering of individuals. In such scenarios, we hypothesize that we may be able to achieve better sample efficiency by focusing on optimizing the ranking of the CATE estimates, as defined by maximizing expected benefit, instead of their accuracy. In **Chapter 4**, we study this mismatch between past work which focuses on optimizing for CATE estimation accuracy and the ultimate goal of optimizing for ranking to assist in informing treatment allocation in the context of constrained resource allocation. We demonstrate that optimizing for ranking may be an easier task than optimizing for accuracy in certain settings, and that better CATE accuracy may not necessarily align with better rankings. Guided by these insights, we propose an approach that directly optimizes for rankings of individuals. Our tree-based approach maximizes the expected benefit of the treatment assignment using a novel splitting criteria. Across synthetic datasets, our approach leads to better treatment assignments compared to CATE estimation methods as measured by expected benefit.

In situations where exact CATE estimation is necessary, there exists a gap between theory

in causal effect estimation and practice. Estimating treatment effects from observational data is challenging, as non-random treatment assignments can lead to confounded and biased estimates. While there exist many popular approaches for estimating CATEs [145, 214], they typically fall into one of three categories. The first category relies only on models of the outcomes (e.g., T-Learner, X-Learner) [112], the second category relies on a model of the treatment assignment (e.e., inverse propensity score weighting) [12, 15], and the third category adjusts for both estimates of the propensity score and the outcome (e.g., R-Learner, DR-Learner) [142, 105]. Approaches that incorporate estimates of both the propensity score and potential outcomes have gained traction in part due to their strong theoretical guarantees in the asymptotic setting, in which models can recover from errors in either the propensity score or the potential outcomes. However, there has been little empirical investigation into how this theory holds in practical settings such as in different levels of confounding or different levels of error in the propensity score. In practice, it remains difficult to select among the multitude of CATE estimation techniques. Moreover, comparisons among techniques are often confounded by differences in the base learning models used. In **Chapter 5**, we explore the gap between theory and practice and present an in-depth exploration of popular CATE estimation techniques. We find that techniques that rely only on estimates of the outcome, in particular the X-Learner, can consistently outperform popular propensity score adjustment techniques.

Finally, we consider how mismatches between ML model development and clinical needs emerge in existing real-world tools through a case study of sepsis risk stratification. Sepsis remains a leading cause of death in hospitals around the world [157, 130, 204, 175]. Timely interventions for individuals diagnosed with sepsis can help reduce downstream mortality, presenting an opportunity for ML models to augment clinical care [188, 53, 117, 160]. Past work has considered building risk stratification and resource allocation models for sepsis based on the likelihood of sepsis infection [208, 44]. However, how these models are trained and evaluated does not match the ultimate goal of augmenting the clinical workflow to improve patient outcomes. In **Chapter 6**, we study this mismatch separately in model development and evaluation. First, existing risk stratification tools are evaluated using predictions made before the time of sepsis. However, sepsis may be clinically recognized and treated before the sepsis definition is met. Predictions occurring after sepsis is clinically recognized may be of limited utility. Prior work has not investigated the accuracy of sepsis risk predictions made before treatment. Thus, we evaluate the discriminative performance of sepsis predictions made throughout a hospitalization relative to the time of treatment. Empirically, we find that a popular sepsis risk stratification model performs no better than random for predicting sepsis when excluding predictions after clinical recognition. Second,

we take a step back and study the mismatch between what existing risk stratification tools optimize for and what is needed to improve patient outcomes in practice. Existing models focus on predicting the risk of sepsis and ignore the potentially heterogeneous effects of the disease. When the likelihood of developing sepsis does not correlate with the effect of sepsis on downstream mortality, targeting those at high risk of developing sepsis may be sub-optimal. To probe the potential shortcomings of this approach, we aim to characterize the heterogeneity of the effect of sepsis on mortality. Across two large clinical populations, we find that there is substantial heterogeneity in the effect of sepsis on the risk of mortality. Moreover, the effect of sepsis on downstream mortality does not strongly correlate with the risk of developing sepsis. Overall, our work explores important gaps between existing sepsis risk stratification tools and the needs of clinical users to improve patient outcomes.

1.2 Contributions

To address the disconnect between ML research for risk prediction and resource allocation and clinical practice, we present several contributions in this dissertation that are summarized below:

- **Calibrated Deep Survival Analysis.** In Chapter 3, we present a new approach for optimizing deep survival models for good discriminative performance, subject to good calibration. Backed by both theoretical and empirical justification, our proposed approach outperforms state-of-the-art techniques and results in more calibrated survival estimates [101].

- **Optimizing for Treatment Allocation to Maximize Expected Benefit.** In Chapter 4, we study the problem of learning resource allocation models to maximize benefit via ranking. We demonstrate the mismatch between accurate CATE estimation and accurate ranking of individuals for maximum benefit. From here, we explore the potential for optimizing for rankings of individuals to inform treatment strategies that maximizes benefit across all treatment thresholds in terms of CATEs compared to baseline techniques that optimize for accurate CATE estimates.

- **Empirically Exploring Mismatches Between Theory and Practice for CATE Estimation.** In Chapter 5, we consider an extensive exploration of CATE estimation techniques in the context of deep learning to understand the mismatch between theoretical results and practice. We explore the performance of popular methods across a variety of relevant settings and highlight key considerations for CATE estimation in practice.

- **Studying the Mismatch Between Sepsis Risk Stratification and Clinical Needs.** In Chapter 6, we study the clinical problem of sepsis risk stratification and identify gaps that may preclude the impact of existing tools in practice. We explore the mismatch between 1) evaluating sepsis prediction models before sepsis time and the goal of alerting clinicians of an individual who may have sepsis before being recognized clinically and 2) predicting the likelihood of sepsis and the ultimate goal of reducing patient mortality.

The promise for ML techniques to reach their potential and truly impact clinical care remains an exciting opportunity [134, 61]. To do this, however, requires an in-depth exploration of the progress made by recent work towards understanding the gap between research and the needs for practical applications in healthcare. In this dissertation, we focus on studying the mismatch between what ML models are optimized and evaluated for in survival analysis and causal effect estimation, and what is needed in certain clinical contexts for risk prediction and resource allocation

The rest of the dissertation is organized as follows. In Chapter 2, we describe relevant background concepts used throughout the remainder of the dissertation. Chapters 3, 4, 5, and 6 describe the technical details of the contributions of this dissertation. Finally, Chapter 7 reflects on future directions related to the work in this dissertation.

CHAPTER 2

Background

In this chapter, we cover important topics referenced throughout this dissertation in the fields of survival analysis and causal effect estimation.

2.1 Survival Analysis

Survival analysis, also known as time-to-event analysis, is a sub-field of statistics focused on learning both the time to some pre-specified event, as well as the probabilistic uncertainty of the event occurring at each time over some time horizon [133]. Despite the name, the application of survival analysis techniques spans many fields, including healthcare, econometrics, finance, and meteorology [64, 25]. In healthcare, survival analysis techniques are not limited to just predicting the onset of death for an individual, as these techniques can also be used to study the time to the occurrence of some adverse event, such as infection, as well [58]. In the context of healthcare, survival analysis techniques can inform clinicians of an individual's or a population's probability of survival over some time interval, allowing said clinicians to properly allocate resources.

Early works in survival analysis focused on learning the distribution of an adverse event over time for a full population [103]. Due to the development and advancement of data acquisition techniques, there has been a rapid rise in the application of ML towards building new survival analysis techniques to learn at an individual-level [202, 73]. These techniques can provide personalized recommendations at a patient level when applied in a healthcare setting.

In the remainder of this section, we first introduce the problem setup and notation used throughout the rest of this dissertation. We then describe recent work in survival analysis, with a particular focus on deep learning techniques developed for the problem. From here, we discuss typical methods for evaluating survival analysis methods.

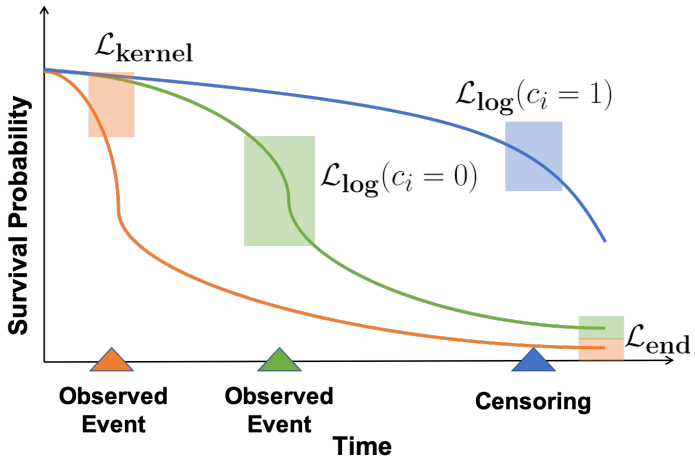

Figure 2.1: Each loss function provides a different kind of supervision throughout the time horizon (shaded region), but none explicitly focuses on calibration.

2.1.1 Problem Set-Up and Notation

Survival analysis aims to learn a time-to-event model using data of the form $D = \{(\mathbf{x}_i, z_i, c_i)\}_{i=1}^n$, where n is the total number of individuals. Each $(\mathbf{x}_i, z_i, c_i) \in D$ represents information for one individual, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the individual's covariates, z_i denotes the observed time of the event, or time of censoring, and c_i denotes the individual's censoring status. Censoring pertains to examples that do not experience any event during the data collection period, and whose event outcomes become unobservable at a certain time. Working with censored data is a primary challenge in survival analysis and remains an important characteristic of the data survival techniques must overcome. In this dissertation, we only consider right-censoring, the most common scenario in survival analysis [41, 103, 187, 202]. An individual i is said to be right-censored ($c_i = 1$) if the event did not occur at time z_i, but instead, the individual was lost to follow-up (*i.e.*, censored) after this time.

In the rest of this dissertation, we focus on accurately estimating individualized survival probabilities over some discrete time horizon [73, 116]. Given data from D, our goal is to learn a model f that maps covariates for individual i $\mathbf{x}_i$ to *individualized* estimates of $P(Z = t|\mathbf{x}_i)$ for $t \in \{0, 1, ..., \tau\}$, where time is binned into τ intervals [116, 156]. From these estimates, we can estimate the survival curves $S(t|\mathbf{x}_i) = P(Z > t|\mathbf{x}_i) = \sum_{j>t} P(Z = j|\mathbf{x}_i)$ and the cumulative incidence function (CIF) $F(t|\mathbf{x}_i) = P(t \leq Z|\mathbf{x}_i) = \sum_{j \leq t} P(Z = j|x_i)$

2.1.2 Deep Survival Analysis

In recent years, researchers have focused on utilizing deep learning when developing survival analysis techniques. Though some works focus on extending traditional techniques, such as the Cox model, to deep learning [104], others have leveraged the flexible nature of deep learning to directly model an individual's underlying survival curve [116, 156]. When doing so, the objective function used to train the model dictates the resulting characteristics of the estimated survival curves. Common objective functions include:

- $\mathcal{L}_{log} = -\sum_{i=1}^{n}(1-c_i)\cdot\log(\hat{P}(Z=z_i|\mathbf{x}_i)) + c_i\cdot\log(\hat{S}(z_i|\mathbf{x}_i))$
- $\mathcal{L}_{end} = -\sum_{i=1}^{n}(1-c_i)\cdot\log(1-\hat{S}(\tau|\mathbf{x}_i))$
- $\mathcal{L}_{kernel} = \sum_{i\neq j} A_{i,j}\cdot exp(\frac{-(\hat{S}(z_i|\mathbf{x}_j)-\hat{S}(z_i|\mathbf{x}_i))}{\sigma})$, where $A_{i,j} = 1_{c_i=c_j=0, z_i<z_j}$

$\mathcal{L}_{log}$, often termed the logarithmic loss, maximizes the estimated probability of the event occurring at the time of observation, while maximizing the estimated survival probability at the time of censoring for censored individuals [116, 156]. $\mathcal{L}_{end}$, often used in conjunction with $\mathcal{L}_{log}$, adds supervision after the observed event time, by forcing the survival probability to zero at the final timestep for uncensored individuals [156]. Lastly, $\mathcal{L}_{kernel}$ penalizes incorrectly ordering two uncensored individuals [116]. **Figure 2.1** shows where these different loss functions provide supervision over the time horizon. Most deep survival models use $\mathcal{L}_{log}$ during training [135, 116, 156]. $\mathcal{L}_{kernel}$ was explored in early deep survival analysis works as a method for increasing ranking performance, but has been less explored recently [116]. DeepHit, a popular feed-forward neural network survival analysis technique, trains its architecture using a composite of $\mathcal{L}_{log}$ and $\mathcal{L}_{kernel}$ [116]. Meanwhile, deep recurrent survival analysis (DRSA), which utilizes a long short-term memory (LSTM) network, has achieved state-of-the-art performance when training using a composite of $\mathcal{L}_{log}$ and $\mathcal{L}_{end}$ [156].

2.1.3 Evaluation

Past work in deep survival analysis often focuses on optimizing and evaluating for discriminative performance. Achieving good discriminative performance means accurately ranking at-risk individuals. Formally, for any two individuals with covariates $\mathbf{x}_1$ and $\mathbf{x}_2$, assume individual 1 has the event at time z_1, at which individual 2 has not had the event nor have they been censored (*i.e.*, $z_2 > z_1, c_1 = 0, c_2 \in \{0,1\}$). Then, we would expect individual 1 to be at greater risk than individual 2 at time z_1, or $\hat{F}(z_1|\mathbf{x}_1) > \hat{F}(z_1|\mathbf{x}_2)$. This is often measured through the C-index, which calculates the proportion of unique pairs of individuals (that match the criteria above) for which this ranking is correct [9, 116].

Another important aspect of survival models towards their clinical applicability is calibration. Well-calibrated models should produce survival estimates $\hat{S}(\cdot|\mathbf{x}_i)$ that match the underlying survival distribution $S(\cdot|\mathbf{x}_i)$. The Brier score, defined at time t as $\frac{1}{n}\sum_{i=1}^{n}(1_{t\leq z_i} - \hat{S}(t|\mathbf{x}_i))^2$, is often used to measure calibration [137, 115, 113]. However, the Brier score measures how well a prediction matches the observed outcome for different individuals, which differs from the definition of calibration considered here. In particular, a discontinuous Heaviside step function that equals 0 at and after the observed event time could qualify as perfectly calibrated as it perfectly matches the observed outcome (*i.e.*, average Brier score = 0), despite no meaningful probabilistic interpretation (i.e., it does not correctly reflect the variation in the probability estimate due to stochasticity in nature). Moreover, the Brier score over the full survival curve is heavily influenced by the choice of time horizon. In **Chapter 3**, we describe and extend metrics more suitable for our definition of calibration [7, 73].

2.2 Causal Effect Estimation

The ability to estimate the effect of an intervention on an outcome for a population, known as the average treatment effect (ATE), has been studied extensively in the past few decades [169, 165, 78]. In particular, information regarding the ATE can help guide clinical care and treatment allocation at a population level. With the increased availability of large observational datasets, recent work has focused on leveraging ML techniques to estimate conditional average treatment effects (CATEs), the effect of a treatment on an outcome, given an individual's covariates [63, 4, 182, 200, 78, 75, 220, 219]. Such effect estimates can guide decision-making in many fields. For example, in healthcare, modeling the effect of an intervention at the individual level can assist clinicians in matching patients to treatments [57]. However, accurately estimating CATEs from observational data is often challenging due to confounding when one or more variables affect both the treatment assignment and the outcome. This leads to fundamental differences between the treatment and control groups, which in turn can lead to inaccurate estimates of the unobserved outcome (i.e., the potential outcome had an individual received a different treatment). Specialized techniques in ML have focused on overcoming this challenge, with a particular focus on deep learning approaches with unique objective functions [182, 219, 220, 90, 74, 75, 185]. In the remainder of this section, we will formally describe the problem of CATE estimation, and introduce some important notation. We will then give some relevant background regarding techniques for unbiased CATE estimation, and how models are generally evaluated.

2.2.1 Problem Set-Up and Notation

We aim to estimate treatment effects given an observational dataset $\mathcal{D} = \{\mathbf{x}_i, t_i, y_i\}_{i=1}^n$, containing n individuals, where each individual i has covariates $\mathbf{x}_i \in \mathbb{R}^d$, has assigned treatment $t_i \in \{0, 1\}$, and experiences the observed outcome under the assigned treatment $y_i \in \mathbb{R}$ (for continuous outcomes) or $y_i \in \{0, 1\}$ (for binary outcomes). We follow the potential outcome framework [169, 189]. Specifically, for an individual i with covariates $\mathbf{x}_i$, we define potential outcomes as the outcomes under different treatment choices (i.e., treated and not treated), and use $Y_i(0)$, $Y_i(1)$ to denote the potential outcomes under non-treatment and treatment respectively. Under the rules of do-calculus, $\mathbb{E}[y|\mathbf{x}_i, do(t = 1)]$ corresponds to the potential outcome $Y_i(1)$ [150]. We define the CATE as:

$$\begin{aligned}\tau_i = CATE(\mathbf{x}_i) &= \mathbb{E}[y|\mathbf{x}_i, do(t = 1)] - \mathbb{E}[y|\mathbf{x}_i, do(t = 0)] \\ &= Y_i(1) - Y_i(0).\end{aligned}$$

The ATE could then be estimated as the average CATE over a population.

2.2.2 Assumptions for Identifiable CATE Estimation

The fundamental problem of causal effect estimation is that we only observe one potential outcome, which corresponds to the observed treatment, and hence we cannot directly estimate the CATE without making some assumptions about the observed data. We follow the vast majority of work in causal inference in making the following assumptions, which are sufficient for the identifiability of the causal effect [84, 78].

Assumption 1 *(No Hidden Confounders). Given $\mathbf{x}_i$, the potential outcomes are independent of the treatment assignment, i.e. $(Y_i(1), Y_i(0)) \perp\!\!\!\perp t_i|\mathbf{x}_i$.* Such an assumption is more likely to hold when the collected covariates are high dimensional, capturing as much information as possible regarding both the treatment assignment and the outcome.

Assumption 2 *(Overlap). Every individual has a non-zero probability of being treated or not, i.e., $0 < P(t|\mathbf{x} < 1), \forall \mathbf{x} \in \mathbb{R}^d, t \in \{0, 1\}$.* The overlap assumption ensures that there are no individuals for whom the treatment assignment is deterministic.

Assumption 3 *(Consistency). The potential outcome of the observed treatment t_i is equal to the observed outcome, i.e. $Y_i(t_i) = y_i$.* The consistency assumption ensures that treatment selection does not change the observed outcome. Hence, under the consistency assumption, we have accurately observed exactly one potential outcome.

2.2.3 Methods for Causal Effect Estimation

In the past years, research in the field of CATE estimation has broadly been focused on a few key areas, including tree-based techniques [79, 13], meta-learning approaches [112, 142, 105], and deep learning approaches [182, 219, 220, 90, 74, 75, 185]. The goal of all techniques is to overcome issues due to confounding, which makes the treatment and control group look dissimilar from one another. Popular methods can be broken down into three main categories: outcome-based models, propensity score adjustment models, and models that adjust using both the propensity score and outcome estimates.

1. Outcome-Based Models. In the first category, algorithms may simply model the potential outcomes using the input covariates in an indirect approach. These estimates of the potential outcomes are used in a second stage of training. Two early techniques in the outcome-based framework were the S-Learner and the T-Learner [112]. The S-Learner appends the observed treatment assignment with the input covariates of the model to learn the observed outcome, while the T-Learner trains two separate models for each treatment group to learn the observed outcomes. In recent years, researchers have proposed a number of CATE estimation techniques that build on the basic ideas of these estimators, with the goal of improving the trade-off for sharing information between treatment groups.

2. Adjustment Using Only Propensity Scores During Training. A second category of techniques adjusts for the propensity score during training. The propensity score, e_i, is the probability that individual i will receive the treatment, i.e., $e_i = p(t = 1|\mathbf{x}_i)$. Adjusting for the propensity score is a sufficient statistic for blocking back-door paths between the treatment and potential outcomes [165, 78]. By reducing the full set of confounders to a single value to control on, propensity score techniques can assist in balancing the distribution of covariates in the treatment groups during training [165, 15, 22]. Traditional propensity score methods often rely on inversely weighting individuals based on estimates of their propensity scores to synthesize a population in which the distribution of covariates is independent of treatment assignment or matching individuals in one treatment group to similar individuals in the opposite treatment group, where similarity is defined using propensity scores [165, 83, 78].

3. Adjustment Using Propensity Score and Outcome Estimates During Training. Finally, the third category of techniques adjusts for both an estimate of the propensity score as well as an estimate of the outcome. These techniques utilize an estimate of both the propensity score and the outcome, often known as "nuisance estimates", to recover if there are errors in one of the models. For example, these methods often build pseudo-outcomes using estimates of the propensity score and the potential outcomes that are unbiased or are efficient estimators regardless of errors in the propensity scores or potential outcomes.

All three categories of models produce CATE estimators that are asymptotically unbiased. However, researchers often prefer methods that incorporate the propensity score over techniques that only model the outcome, especially in high-dimensional settings with large levels of confounding [30]. In these situations, learning models that transfer between treatment groups using only outcome-based models may be difficult without explicitly accounting for this confounding bias, as shown by past theory [3, 22]. Propensity score adjustment techniques are singly robust, in that they only guarantee unbiased CATEs if the propensity score is accurate. Methods that rely on estimates of both the propensity score and the potential outcomes provide stronger theoretical robustness guarantees to errors in the propensity score. For example, these methods show that errors in the propensity score can be mitigated with a strong estimate of the outcomes, or that despite errors in the propensity score, we may still achieve oracle efficiency (i.e., efficiency assuming perfect propensity scores and potential outcomes) for CATE estimation [142, 105]. Given their strong theoretical basis, methods incorporating both estimates may be preferred for CATE estimation.

2.2.4 Evaluation

The fundamental challenge of causal effect estimation prohibits direct evaluation on real-world data in which ground-truth treatment effects are not available. Instead, past works often utilize fully synthetic or semi-synthetic datasets, in which ground-truth treatment effects and confounding can be simulated. Past works consider calculating the ground-truth performance of each model in terms of the precision in estimating heterogeneous treatment effects (**PEHE**), defined as

$$\sqrt{\epsilon_{PEHE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tau_i - (\hat{Y}_i(1) - \hat{Y}_i(0)))^2}$$

[182, 74, 75, 92]. A smaller value of ϵ_{PEHE} represents more accurate CATE estimates. Past work has often measured the value of a CATE estimator based on the resulting models ϵ_{PEHE} across benchmark datasets.

CHAPTER 3

Calibrated Deep Survival Analysis

3.1 Introduction

To begin, we consider the task of predicting an individual's risk over time by studying problems in the field of survival analysis. In survival analysis, one aims to learn the relationship between an individual's covariates and the underlying stochastic process of some event (*e.g.*, disease onset). Beyond discriminative performance (*i.e.*, how the relative predictions between individuals match the observed outcomes), to be useful for real-world applications, survival models must be well calibrated [65]. In clinical settings, making decisions at a patient-level requires survival estimates that are accurate with respect to the ground-truth survival probability. Poor calibration can lead to misleading predictions, resulting in potentially clinically harmful models [197, 196, 180, 191]. Accurate estimates of survival at different time-points can help augment clinical decision-making at a per-patient level.

We define a calibrated model as one that consistently produces estimates of survival that match the underlying survival probabilities for each individual [73]. To better understand what these individualized underlying survival probabilities represent, consider building an estimate of survival for an entire population using a simple counting-based Kaplan-Meier estimate [103]. Differences among individuals will lead to events at different time-points, resulting in a decreasing estimate of population-level survival over time. This estimate reflects the variation in the time-to-event distribution. Now consider a set of individuals with identical or near-identical covariates. Despite the similarity among individuals, we might still expect the time-to-event distribution for this homogeneous population to exhibit some variation due to the stochasticity in nature, resulting in a gradually decreasing survival curve for these individuals. Along these lines, the underlying survival curve for an individual should reflect such variation.

Figure 3.1 illustrates potential differences in discriminative performance and calibration performance via a hypothetical example. The *solid* curves represent the true underlying

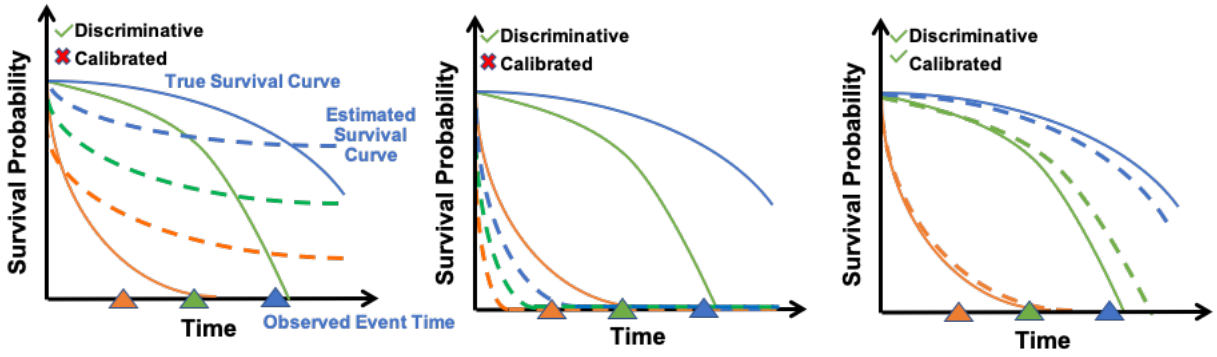

Figure 3.1: Hypothetical Example. Three hypothetical sets of estimated survival curves for three individuals (dashed) and their corresponding true underlying survival distributions (solid), where the triangles represent the observed event times. All three sets of estimated curves correctly rank the individuals (*i.e.*, have good discriminative performance). However, the first two sets of estimated survival curves consistently overestimate or underestimate the true survival probability at various points throughout the time horizon. Meanwhile, the third set of estimated survival curves closely aligns with the true survival curves. Hence, the estimated survival curves more accurately reflect the probability of survival. The first two sets of estimated survival curves are *miscalibrated*, while this third set of estimated survival curves is *well-calibrated*.

survival distributions, and the *dashed* lines represent hypothetical estimates for three different individuals. With respect to the observed event times, all three sets of estimated survival curves *correctly rank* the individuals, and hence, have good *discriminative* performance. However, the first two sets of survival curves (a and b) consistently underestimate or overestimate the survival probabilities with respect to the true survival curve. Hence, these estimates are *miscalibrated*. Meanwhile, the third set of estimated survival curves (c) is well calibrated, since it aligns with the true survival probabilities. These calibrated estimates provide an accurate probabilistic interpretation of survival for an individual throughout the time horizon.

Deep survival models have achieved state-of-the-art discriminative performance by relaxing any distributional assumptions and directly estimating the underlying process [116, 156]. However, to date, such models are trained by optimizing for discriminative performance and have not been evaluated in terms of calibration. Though useful for ranking individuals, the resulting survival curves may consistently overestimate or underestimate an individual's probability of survival, as in **Figure 3.1**.

In light of these issues, we focus on approaches for training and evaluating deep survival analysis models that account for *both* calibration and discriminative performance. In this chapter, our contributions include:

- we highlight the shortcomings of existing methods for training and evaluating in terms of calibration,
- we propose a novel training scheme for deep survival analysis models and provide theoretical justification for why this training scheme should result in well-calibrated survival estimates, and
- we empirically demonstrate that the proposed training scheme leads to well-calibrated models while remaining competitive in terms of discriminative performance

We present a framework for training and evaluating deep survival models that focuses on calibration. Through a series of experiments on two publicly available datasets, we compare our approach to state-of-the-art approaches in survival analysis, demonstrating the proposed approach's ability to maximize discrimination subject to good calibration.

3.2 Background and Related Work

In **Section 2.1**, we describe the problem set-up for survival analysis and current techniques for training deep survival models, including the logarithmic loss. Though the logarithmic loss corresponds to a proper scoring rule, it is sensitive to extreme cases and outliers [66, 65]. This sensitivity results in a larger trade-off between making accurate predictions and maintaining calibration compared to other proper scoring rules, such as the continuous-rank probability score (CRPS). These methods have not been evaluated for their calibration performance. We hypothesize that the models trained to minimize $\mathcal{L}_{log}$ could result in miscalibrated survival estimates. In light of this observation, we consider loss functions that build off of proper scoring rules without this limitation. In particular, our proposed approach builds on the CRPS, which is defined as $\int_{-\infty}^{\infty}(\hat{F}(t|\mathbf{x}_i) - 1_{z\leq t})^2 dt$, which has been explored in survival analysis [17]. However, this objective function relies on an infinite integral and thus requires specific distributional assumptions during training. In contrast, our discrete approximation avoids relying on any distributional assumptions. Moreover, we consider how this discrete approximation can be incorporated into a training scheme with other loss functions to elicit calibrated and accurate survival estimates. Finally, we consider a comprehensive evaluation framework for properly measuring the efficacy of survival models for both their discriminative performance and calibration. Concurrent work to ours proposed directly optimizing for a variant of a calibration metric we use for evaluation [67]. Future work might consider how the two proposed training schemes could be combined for further improvements.

3.3 Methods

In this section, we present our proposed training scheme and our comprehensive evaluation metrics. We begin by proposing a new loss function and theoretically justifying why it should elicit survival models with good discriminative performance and good calibration. We continue by discussing and justifying our proposed training scheme, which consists of combining this new loss function with $\mathcal{L}_{kernel}$. We explain why this combination should improve both overall performance. We conclude with a discussion on how to evaluate models for both discriminative performance and calibration.

3.3.1 Proposed Training Scheme

We propose minimizing the rank probability score (RPS), $\mathcal{L}_{RPS}$, defined as:

$$\sum_{i=1}^{n}(1-c_i)\cdot\sum_{t=1}^{\tau}(\hat{S}(t|\mathbf{x}_i)-1_{t<z_i})^2+c_i\cdot\sum_{t=1}^{z_i}(\hat{S}(t|\mathbf{x}_i)-1)^2$$

$\mathcal{L}_{RPS}$ focuses on the relevant portions of the full time horizon, rather than just the specific event-time. For uncensored individuals ($c_i = 0$), $\mathcal{L}_{RPS}$ pushes the survival probability at times before an individual has an event to 1, and shrinks the survival probability to 0 at times after the event has occurred. For uncensored individuals, $\mathcal{L}_{RPS}$ is averaged over the full time horizon τ, as we have access to the survival status for the full time interval. For censored individuals ($c_i = 1$), $\mathcal{L}_{RPS}$ pushes the survival probability to 1 before the individual is censored, and is averaged over the available time horizon for censored individuals z_i, as we do not know their survival status after this time.

Claim. *Training deep survival models using $\mathcal{L}_{RPS}$ will result in well-calibrated estimates of survival.*

Proof. Consider n individuals with identical or near-identical covariates with observed event times $\{z_i\}_{i=1}^n$. Define the counting-based Kaplain-Meier estimate for these individuals at time t as $KM_t^n = \frac{1}{n}\sum_{i=1}^n 1_{t<z_i}$, where $\lim_{n\to\infty} KM_t^n$ is the underlying survival probability at time t for these n individuals.

A survival model will estimate one survival probability for these n individuals at time t. Define this value as $\hat{p}_t$. A well-calibrated survival model will output a $\hat{p}_t$ that closely aligns with the underlying survival probability $\lim_{n\to\infty} KM_t^n$. Consider the optimization problem of finding $\hat{p}_t$ which will minimize $\mathcal{L}_{RPS}$. This problem can formally be set-up as $\arg\min_{\hat{p}_t}\sum_{i=1}^n(\hat{p}_t - 1_{t<z_i})^2$.

First, this optimization problem is strictly convex and has a unique minimum, as the second derivative is positive everywhere (see **Appendix A.2** for more detail).

To find the value of $\hat{p}_t$ that minimizes this objective function ($\hat{p}_t^*$), we set the derivative equal to zero.

$$\frac{\partial}{\partial \hat{p}_t^*}\left(\sum_{i=1}^{n}(\hat{p}_t^* - 1_{t<z_i})^2\right) = 0$$

$$2\hat{p}_t^* - \frac{2}{n}\sum_{i=1}^{n} 1_{t<z_i} = 0$$

$$\hat{p}_t^* = \frac{1}{n}\sum_{i=1}^{n} 1_{t<z_i}$$

The unique estimated survival probability that minimizes the objective function is equivalent to the average survival status for all n individuals at time t. This unique minimum is equal to KM_t^n which, as n gets large, is equal to the true underlying survival probability for these individuals at time t. Hence, training a survival model to minimize $\mathcal{L}_{RPS}$ will result in estimated survival probabilities that align well with the true survival probabilities. □

A model that minimizes $\mathcal{L}_{RPS}$ will theoretically result in well-calibrated survival estimates that align well with the true survival curves. However, due to the inherent noise in the training process of deep models and the inability to guarantee a global solution, training using just $\mathcal{L}_{RPS}$ as a loss function might be insufficient. In particular, combining $\mathcal{L}_{RPS}$ with a loss function that can scale survival probabilities and encourages good discriminative ability would improve overall performance.

Hypothesis. *Training deep survival models using a composite loss function $\mathcal{L}_{RPS} + \lambda\mathcal{L}_{kernel}$, yields an accurate, yet calibrated survival model when the value of σ in $\mathcal{L}_{kernel}$ is appropriately tuned.*

Justification. Remember that $\mathcal{L}_{kernel}$ is defined as $\mathcal{L}_{kernel} = \sum_{i\neq j} A_{i,j} \cdot exp(\frac{-(\hat{S}(z_i|\mathbf{x}_j) - \hat{S}(z_i|\mathbf{x}_i))}{\sigma})$. In this loss function, σ controls the scale of the differences between survival probabilities for different individuals. When σ is small (*i.e.* $\sigma \leq .1$) and individuals are correctly ranked, small or large differences between two individual's survival probabilities (numerator) minimize $\mathcal{L}_{kernel}$. In contrast, when σ is large (*i.e.* $\sigma \geq 10$) and individuals are correctly ranked, only large differences between individual's survival probabilities can minimize $\mathcal{L}_{kernel}$. Hence, the value of σ can directly affect how the variation of different individual's survival curves over the interval $[0, 1]$. In particular, we expect that training a model to minimize $\mathcal{L}_{kernel}$ with a small σ value will result in survival curves that are not well-spread out while training a model to minimize $\mathcal{L}_{kernel}$ with a large σ value will scale the survival curves in order to spread them out sufficiently. The value of σ should be tuned based on a validation set.

The ability to control the variation of individuals' survival curves can also be thought

of as rescaling survival curves in order to best minimize $\mathcal{L}_{kernel}$. If $\mathcal{L}_{RPS}$ overestimates or underestimates the survival probability for individuals at certain times, using $\mathcal{L}_{kernel}$ with an appropriately tuned value of σ can scale these estimates to more accurately estimate the true underlying survival probabilities. At the same time, as $\mathcal{L}_{kernel}$ aims to correctly rank individuals, it will still maximize discriminative performance. Thus, we expect that the combination of $\mathcal{L}_{kernel}$ and $\mathcal{L}_{RPS}$ will encourage good calibration without sacrificing discrimination.

The value of λ helps control the trade-off between the two loss functions in the composite loss. As setting λ to 0 translates to simply the $\mathcal{L}_{RPS}$ loss function, and setting λ too high translates to the $\mathcal{L}_{kernel}$ loss function, we hypothesize that an intermediate value of λ will result in the best trade-off between the theoretical guarantees of correctly estimating the underlying survival probability obtained by minimizing $\mathcal{L}_{RPS}$ and the scaling ability of $\mathcal{L}_{kernel}$.

In summary, we introduced a novel loss function $\mathcal{L}_{RPS}$, which we hypothesize will result in increased calibration performance when used to train survival analysis models. Moreover, we proposed a new training scheme that involves minimizing a composite loss of $\mathcal{L}_{RPS}$ and $\mathcal{L}_{kernel}$.

3.3.2 Evaluating Model Performance

We evaluate model performance in terms of both discrimination and calibration. We evaluate discriminative performance, in terms of the aforementioned C-index, which calculates the proportion of individuals who are correctly ranked by the estimated models. To measure calibration, we consider the average Brier score (*i.e.*, mean-square-error over the survival curve) and D-Calibration [73, 7]. Brier score measures how well a prediction matches the observed outcome for different individuals, and hence, does not fully capture our definition of calibration. D-Calibration bins the estimated survival probabilities at the true event times into ten equal-width intervals between 0 and 1 and performs a chi-squared test to determine if the distribution is uniform. This more closely aligns with our definition of calibration; however, the test assumes the model is well-calibrated, placing the burden on disproving the null hypothesis.

In light of these shortcomings, we also consider the **distributional divergence for calibration (DDC)**. DDC does not rely on a statistical test and produces a continuous score that allows for comparisons of different models. Given a set of estimated survival probabilities for each individual at their observed event times $\{\hat{S}(z_i|\mathbf{x}_i)\}_{i=1}^n$, we compute DDC as the Kullback-Leibler (KL) Divergence $D_{KL}(P||Q)$ between a binned distribution

$P = B(\{\hat{S}(z_i|\mathbf{x}_i)\}_{i=1}^n)$ and the uniform distribution Q, where B is a function that maps a set of probabilities into a probability distribution over $\mathcal{X}$, ten equal-width bins covering the unit interval [119]. Due to the discrete nature of the binning operation, we change the base of the logarithm when calculating DDC to ensure that it ranges between 0 and 1. DDC measures the distance between the empirical distribution of estimated probabilities of survival at the time of the events P and the uniform distribution Q. Lower is better; if $P = Q$, then $DDC(P, Q) = 0$. Survival curves that estimate a single survival probability, such as 0, for every individual at their observed event time, for which $B(\{\hat{S}(z_i|\mathbf{x}_i)\}_{i=1}^n) = B(\{0\}^n)$, achieve a maximum DDC of 1.

Claim. *A perfectly calibrated survival model necessarily minimizes the divergence between P and Q for a sufficiently large n.*

Proof. The probability integral transform states that for some random variable X with cumulative distribution function F_x, $F_x(X)$ should be uniformly distributed $U(0, 1)$ [8]. Thus, given a randomly sampled event time z_i, it must be that $S(z_i) = 1 - F(z_i) \sim U(0, 1)$. Given a set of randomly sampled event times $\{z_i\}_{i=1}^n$, where n is sufficiently large (e.g., $n >> 10$), we then expect the distribution of $P = B(\{\hat{S}(z_i|\mathbf{x}_i)\}_{i=1}^n) \sim U(0, 1)$ [73]. Hence, a calibrated survival model should minimize the divergence between P and a uniform distribution Q. □

Though necessary, minimizing this metric does not *guarantee* that the estimated survival curves accurately estimate the true underlying survival process. Despite good calibration, these probabilistic estimates may still be inaccurate (i.e., poor discrimination). Hence, it is important to evaluate models in terms of both their calibration and their discriminative performance. To this end, we seek models that excel with respect to *both* measures of performance.

Importantly, DDC is not applied to censored individuals. Though learning with censored individuals is a key element of survival analysis, evaluating calibration on censored individuals raises a number of issues. Without strong assumptions on the event time distribution for censored individuals, one cannot make meaningful conclusions regarding the calibration of a model for censored individuals (see **Appendix A.3** for discussion). To this end, while we measure discriminative performance across both uncensored and censored individuals, we focus our evaluation of calibration (specifically, DDC and D-Calibration) on uncensored individuals. This introduces a mismatch between the distribution we evaluate in practice and the one we aim to evaluate in theory. However, if patients are censored at random, this estimate of calibration should generalize.

Tradeoff between calibration and discrimination. It is important to note that well-calibrated survival curves need not have optimal discriminative performance on the observed sample. A well-calibrated model is one that consistently estimates survival curves that closely

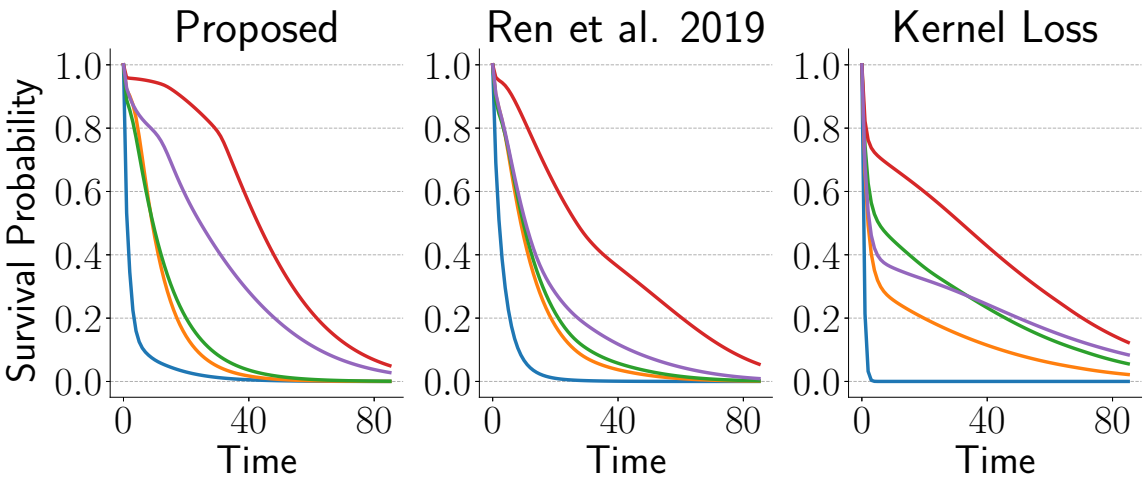

Figure 3.2: Example survival curves estimated using DRSA trained with $\mathcal{L}_{log} + \mathcal{L}_{end}$ (left), example survival curves estimated using DRSA trained with the proposed training scheme (middle), and example survival curves estimated using DRSA trained with $\mathcal{L}_{kernel}$ (right) on the NACD dataset. Each color represents a randomly selected individual from the test set; the same individuals are shown in each graph. Visually, training with the proposed scheme results in survival curves with a greater variation in shape over time, due to the supervision over the full time horizon and the relative scaling abilities of $\mathcal{L}_{kernel}$.

match the true survival curves. Due to stochasticity, some individuals may experience the event when their true survival probability is high. As a perfectly calibrated model will estimate a high survival probability at the observed event time for these individuals, these individuals will contribute negatively to the C-Index calculated based on the observed event times. Hence, when individual time to event varies (which we expect is often the case due to the stochasticity of nature), there exists a trade-off between obtaining perfect calibration and perfect observed discriminative performance (i.e. a C-index of 1). This issue arises due to discrimination being measured with respect to only a single observed sample. This phenomenon is explored further in **Appendix A.4**.

Practically speaking, both measures of performance are important. Maintaining discriminative performance with increased calibration represents an important gain for a particular survival model. Accordingly, we consider the trade-off between the discriminative performance and calibration by calculating the harmonic mean between the C-index and $1-DDC$, a value we term the **total score**. A higher total score corresponds to a model that balances discriminative performance and calibration.

3.4 Experiments and Results

Empirically, we test our hypothesis that the proposed approach will outperform baseline techniques in terms of the trade-off between calibration and discriminative performance. We present two publicly available datasets on which we test our proposed methods and benchmark methods to which we compare. We detail the proposed method's performance compared to the benchmarks in terms of discrimination and calibration and compare against different ablations of the proposed method using the new evaluation framework.

3.4.1 Experimental Setup

Datasets. We consider two publicly available datasets:

- the **Northern Alberta Cancer Dataset (NACD)** consists of 2,402 individuals with various forms of cancer [73, 217]. The dataset tracks 51 features for each individual, including demographics, vital signs, patient characteristics such as appetite, and specific details about the type and progression of the cancer. 36.6% of the individuals in the dataset are right-censored, with an average survival time of 16.06 months for uncensored individuals. For this dataset, we use a τ of 86 months based on the largest length of stay.

- **CLINIC** records the survival status of 6,036 patients in a hospital, with 13.2% being censored [110]. The dataset consists of 14 features for each individual, including information about demographics, vital signs, onset of diseases, and medications. The average survival time for uncensored individuals is 5.33 months. For this dataset, we use a τ of 52 months, such that each time-bin represents one month.

Model Architecture. To demonstrate the efficacy of our approach and compare it against baseline methods, we consider minimizing the proposed composite loss to train the Deep Recurrent Survival Analysis (DRSA) architecture [156]. Though the proposed loss functions are model-agnostic, we consider the DRSA architecture due to its state-of-the-art discriminative performance, flexibility for allowing variable-length forecasting, and lack of assumptions regarding the probability at the end of the time horizon. More information about this architecture choice can be found in **Appendix A.1**.

Baselines. To evaluate how our proposed approach compares to current state-of-the-art in deep survival analysis, we compare against two baseline survival analysis models:

- The DRSA architecture with the objectives it was originally proposed with (using $\mathcal{L}_{log}$ and $\mathcal{L}_{end}$) [156], and

Model	NACD				
	C-index ↑	**DDC ↓**	**D-Calibration ↑**	**$\overline{\text{Brier}}$ ↓**	**Total Score ↑**
Ren et al. 2019	$.748 \pm .002$	$.025 \pm .012$	1	$.101 \pm .002$	$.846 \pm .004$
MTLR	$.750 \pm .000$	$.062 \pm .000$	0	$.101 \pm .000$	$.834 \pm .000$
Proposed - $\mathcal{L}_{RPS}$	$.741 \pm .008$	$.305 \pm .089$	0	$.207 \pm .034$	$.715 \pm .050$
Proposed - $\mathcal{L}_{kernel}$	$.742 \pm .003$	$.012 \pm .002$	3	$.101 \pm .003$	$.847 \pm .001$
Proposed	$.742 \pm .006$	$\mathbf{.007 \pm .003}^*$	5	$.104 \pm .002$	$.850 \pm .003$
Model	**CLINIC**				
	C-index ↑	**DDC ↓**	**D-Calibration ↑**	**$\overline{\text{Brier}}$ ↓**	**Total Score ↑**
Ren et al. 2019	$.616 \pm .003$	$.138 \pm .002$	0	$.107 \pm .000$	$.719 \pm .003$
MTLR	$.608 \pm .000$	$.168 \pm .000$	0	$.106 \pm .000$	$.702 \pm .000$
Proposed - $\mathcal{L}_{RPS}$	$.628 \pm .003$	$.241 \pm .022$	0	$.153 \pm .002$	$.687 \pm .011$
Proposed - $\mathcal{L}_{kernel}$	$.615 \pm .005$	$.097 \pm .006$	0	$.110 \pm .001$	$.731 \pm .005$
Proposed	$.627 \pm .001$	$\mathbf{.056 \pm .011}^*$	0	$.106 \pm .001$	$\mathbf{.753 \pm .004}^*$

Table 3.1: The proposed training approach consistently leads to improvements in calibration (DDC, D-Calibration, Averaged Brier Score) across all baselines and ablations, without sacrificing discriminative performance (C-index) (mean ± standard deviation across random initializations, number of times passing the statistical test for D-Calibration). Lower DDC and Brier scores and higher values of C-index, D-Calibration, and total score indicate better performance. An * indicates results that are statistically significant over all baselines using a paired t-test ($p < .05$).

- Multi-task logistic regression (MTLR) is one of the only survival analysis approaches that has shown good empirical performance in terms of our definition of calibration [217, 73]. MTLR trains a separate logistic regression model per time-point to estimate survival and combines these to estimate the survival distribution over some time horizon. When compared to other methods, such as extensions of the Cox model, MTLR performed best in terms of both calibration and discrimination [73].

Training/Evaluation Details. Across experiments, we use the same DRSA architecture: a one-layer LSTM with hidden size 100 and a single feed-forward layer with a sigmoid activation on the output for each time-step [156]. We separate our data into training/validation/test sets using a 60/20/20% split. For training, we use Adam and a batch size of 50 [107]. We train for 100 epochs (which, empirically, was enough for models to converge) and select the best model based on a validation set. For the proposed composite training scheme, we tune the value of σ for $\mathcal{L}_{kernel}$ based on the NACD dataset, and use this optimal value on the CLINIC dataset to test whether the manner in which $\mathcal{L}_{kernel}$ affects $\mathcal{L}_{RPS}$ generalizes across multiple datasets. When training with multiple losses, we use $\lambda = 1$. Though we considered other weighting schemes, it did not appear to affect performance. Note that

we weighted the $\mathcal{L}_{RPS}$ loss function due to the right-skewed time-to-event distribution for both datasets. We train each model five times, with different weight initializations. We present the mean and the standard deviation of the results on the test set for all metrics except D-Calibration, for which we present the number of runs where the resulting survival estimates passed the D-Calibration test. We evaluate DDC and D-calibration using only uncensored test individuals, but we evaluate C-index and Brier score using all test individuals. All deep models were built in PyTorch [1], while MTLR was implemented using the corresponding R package [149, 72].

3.4.2 Results

First, our proposed approach consistently outperforms all baselines with respect to DDC and D-calibration, while maintaining comparable C-index and average Brier score values (**Table 3.1**). Lower values represent better performance for DDC and Brier score, while higher values represent better performance for the other metrics. The proposed method consistently leads to estimated survival curves with a better trade-off between calibration and discrimination, as evidenced by the higher total score compared to MTLR and DRSA as it was originally proposed. The fact that no model dominates in C-index across datasets is consistent with recent findings in survival analysis [115].

Compared to the original DRSA [156], the proposed training scheme results in a statistically significant improvement in calibration across both tasks (NACD DDC: .025 vs. .007, CLINIC DDC: .138 vs .056). This improvement, however, is accompanied by a small decrease in C-index in the NACD dataset. However, the probabilistic estimates of survival are more likely to accurately represent the true underlying survival processes. We see the same overall trend when comparing our proposed method with MTLR, where the proposed model is significantly more calibrated across both datasets (NACD DDC: .062 vs .007, CLINIC DDC: .168 vs .057), while the relative C-index depends on the dataset.

Compared to training each component of the proposed loss (*i.e.*, $\mathcal{L}_{RPS}$ and $\mathcal{L}_{kernel}$) separately, using the composite loss leads to improvements (**Table 3.1**: NACD total score: .715 and .847 vs .850, CLINIC total score: .687 and .731 vs .753). In particular, note that training with $\mathcal{L}_{RPS}$ results in good calibration performance, while training with $\mathcal{L}_{kernel}$ in and of itself results in poor calibration performance. Hence, as expected, $\mathcal{L}_{RPS}$ itself will elicit calibrated and accurate estimates of survival, but combining it with the scaling ability of $\mathcal{L}_{kernel}$ can improve performance even more. Moreover, training using $\mathcal{L}_{RPS}$ alone results in better calibration than using the logarithmic loss functions (NACD DDC: .025 vs .012, CLINIC

[1] https://github.com/MLD3/Calibrated-Survival-Analysis

DDC: .138 vs .097), with minimal drops in discriminative performance. These empirical results support the original hypothesis that training using $\mathcal{L}_{RPS}$ should result in survival models that better balance discriminative performance and calibration, but the composite loss results in the best performance.

Next, we focus on a qualitative assessment of our proposed method. Visually, this approach produces survival curves with a greater variation in shape over the full time horizon (**Figure 3.2**). In particular, the baseline training scheme results in survival curves that decay quickly towards a survival probability of 0. This is evidenced by the high DDC value due to many individuals' estimated survival probabilities being very low at the time they experienced the event. Meanwhile, our proposed loss functions achieve better DDC values by allowing more flexibility in the shape of the survival curves, such that some individuals have higher survival probabilities at the time of their observed events. We hypothesize that this is due in part to the direct supervision over the entire predictive distribution that comes from training with $\mathcal{L}_{RPS}$. In contrast, $\mathcal{L}_{log}$ provides direct supervision on the survival probability over only a single time-point, possibly resulting in less flexibility in the shape of the predictive distribution over the time horizon [64]. This single time-point supervision, along with the logarithmic losses sensitivity to extreme cases, can result in miscalibrated survival curves.

We present results for the proposed method using $\sigma = 0.8$ in $\mathcal{L}_{kernel}$ for both datasets. This value was tuned on a validation set on the NACD dataset and applied to the CLINIC dataset. Hence, the manner in which $\mathcal{L}_{kernel}$ affects $\mathcal{L}_{RPS}$ generalizes across multiple datasets, supporting our original hypothesis. Moreover, we visually confirm the original motivation for the use of $\mathcal{L}_{kernel}$: the value of σ helps control the scale of different individual's survival curves. As noted in Section 3, we expect a model trained to minimize $\mathcal{L}_{kernel}$ with small σ (*e.g.* $\sigma = 0.1$) to result in survival curves where different individuals curves are close to each other in scale, and a model trained to minimize $\mathcal{L}_{kernel}$ with large σ (*e.g.* $\sigma = 10.0$) to result in more spaced out survival curves. **Figure 3.3** shows estimated example curves for 10 random individuals in the NACD dataset when trained using $\mathcal{L}_{kernel}$ with σs of 0.1 and 10.0. The resulting survival curves display the hypothesized phenomenon, confirming the ability of $\mathcal{L}_{kernel}$ to control the scale of different individuals' survival curves. Hence, the improved performance for the composite loss is in part due to an additional rescaling of the survival distributions to better match the underlying survival probabilities.

Overall, these results indicate the ability of our proposed training procedure to better match the true survival distribution, while maintaining the useful property of accurately ranking individuals. Moreover, the comprehensive evaluation framework helps facilitate model comparisons for both discriminative performance and calibration.

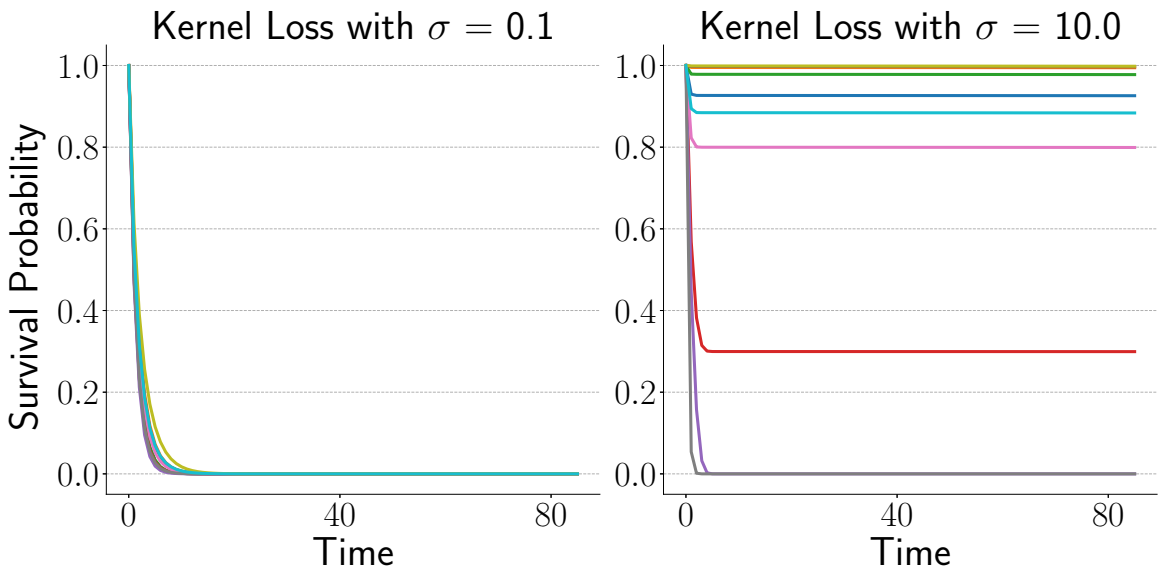

Figure 3.3: Survival curves from models trained with $\mathcal{L}_{kernel}$ using $\sigma = 0.1$ (left) and $\sigma = 10.0$ (right) from the NACD dataset. Each color represents a different individual. These plots confirm our original hypothesis regarding $\mathcal{L}_{kernel}$: the value of σ can control the relative scales of survival probabilities. Hence, by tuning σ, we can scale the survival curves to best match the true underlying survival distributions.

3.5 Conclusion and Discussion

Given the stochasticity of nature, we expect individuals to have an *underlying survival distribution* that corresponds to a meaningful probabilistic interpretation of an individual's survival. Though critical to clinical application, calibration to date has been largely overlooked in survival analysis, especially in deep survival analysis. We hypothesized that recent work in deep survival analysis that optimizes and evaluates for discriminative performance alone results in poorly calibrated estimated survival curves. To this end, we introduced a new approach for training deep survival analysis models to optimize for both discriminative performance and calibration. We provided both theoretical justification and empirical evidence for why the proposed approach elicits calibrated estimates of survival. Applied in the context of a state-of-the-art deep survival analysis architecture, the proposed training scheme leads to significant gains in calibration across two publicly available datasets, while achieving similar discriminative performance. Still, there remains room for improvement. In particular, handling continuous-time survival analysis problems without the use of any distributional assumptions is an interesting line of future work. Nonetheless, this chapter presents a complete and flexible pipeline for training and evaluating accurate and well-calibrated deep models for survival analysis.

CHAPTER 4

Learning to Rank for Treatment Allocation

4.1 Introduction

We next consider the problem of resource allocation or prioritizing interventions, a common problem across across various fields [27, 111, 140, 39]. In healthcare, for instance, clinicians must triage patients for different levels of care [163]. In marketing, companies must prioritize customers for marketing campaigns and retention programs [10, 153]. Similarly, in education, targeted interventions can lower dropout rates or better school performances [18, 146]. While numerous other examples exist, in this section, we continue in using the healthcare setting as a motivating example.

In many healthcare settings, the optimal situation may be to treat all at-risk patients. However, due to resource constraints such as time, workforce, and availability of treatments, healthcare workers often have to make important and difficult decisions on how to allocate resources [108, 68]. For example, clinicians may have to prioritize extra monitoring and care to a subset of individuals at risk of deteriorating due to some disease, such as sepsis. Such decisions are often multi-faceted. However, one aspect of the decision might consider who would benefit most from the decision. Then, the decision could be based on a ranking of who is likely to benefit most from a particular resource or intervention [108, 177, 212, 85]. Tools that could help clinicians in estimating benefit from observational data could help in assisting clinicians in defining this ranking. However, estimating treatment effects is not always straightforward.

Conditional average treatment effects (CATEs) quantify the effect of a treatment on an outcome given an individual's covariates using observational data. However, estimating CATEs is challenging due to confounding when variables affect both the treatment assignment and outcome [57, 78]. Accordingly, past research has worked to improve accuracy and sample efficiency in CATE estimation through novel machine learning techniques [63, 4, 182, 200, 78, 75, 220, 105]. However, these methods are often optimized for and

evaluated based on their ability to *accurately* estimate CATEs.

More recently, there has been interest in how causal inference techniques translate to downstream decision-making. Specifically, researchers have studied when exact causal effect estimation may be unnecessary when the goal is to identify whom to treat and framed a new problem of causal classification for identifying treatment responders [96, 14, 52]. In these settings, the goal is to learn whether an individual will benefit from treatment, as defined by some threshold, and prioritize treatment for these individuals. Past work has both studied the disconnect between this problem and CATE estimation and has studied methods towards directly optimizing for this use case. In this section, we build upon this recent paradigm shift and extend this idea even further beyond a binary classification problem and study the problem of resource allocation policies without the need for an a priori threshold to treat, similar to triage. As these thresholds for determining who to treat may vary depending on the application, and may change many times within the same application, it remains essential to build models agnostic to a particular threshold when everyone benefits from the treatment.

Recent research in the field of uplift modeling, which often assumes access to data from a randomized controlled trial has begun to study this problem [171, 21, 222, 223]. For example, [223] proposes a new loss function to directly obtain unbiased CATE estimates that may be used to rank individuals for resource allocation. While related with regards to the interest in treatment allocation, past work often assumes access to data from a randomized controlled trial or with binary outcomes. This difference in the problem setting changes the problem substantially, such that their proposed estimators, and the theory underlying their estimators, no longer apply as the outcomes and treatments are not independent in our observational setting. Moreover, we focus on studying the disconnect between the problem of optimizing for treatment allocation based on expected benefit and unbiased CATE estimation, which often remains the objective of past work [223]. Finally, recent work from [51] studies how confounded data may affect the task of ranking causal effects, and posit a rank-preserving assumption that would allow an accurate ranking of CATEs even without access to all relevant confounders. In this section, we assume access to all confounders and study the relationship between optimizing for accurate CATE estimation and accurate ranking, with a focus on the potential for directly optimizing for ranking rather than some estimate of CATE towards maximizing benefit in resource allocation. Building on these recent works, we focus on a theoretical and empirical exploration of the disconnect between these two problem set-ups. We focus on a setting in which the treatment may be beneficial to many people, but due to resource constraints, it must be allocated to those who benefit most from the treatment. We take inspiration from the field of learning to rank to tackle this problem

and consider how to adapt these methods to our setting [32].

In the context of resource allocation, accurate CATE estimates will produce an accurate ordering of who is most likely to benefit from the resources. While sufficient, accuracy in CATE estimation is not necessary. Inaccurate or biased estimates can still lead to the best ranking, i.e., one that maximizes benefit across all treatment thresholds. In this section, we study the disconnect between accurate CATE estimation and the ultimate goal of prioritization for resource allocation. We theoretically analyze the mismatch between optimizing for CATE estimation accuracy and optimizing for rankings towards maximizing downstream benefit. Based on our findings, we develop a novel approach that aims to learn an accurate resource allocation ranking. We propose a tree-based approach that produces a ranking of individuals that maximizes expected benefit across all treatment thresholds. We show that our approach focused on optimizing for ranking and benefit is more sample-efficient and outperforms CATE estimation techniques that focus on accuracy in low-sample settings. Overall, our contributions are as follows:

- We analyze the problem of learning accurate ranking models for maximum benefit compared to learning accurate CATE estimation models.
- We propose a novel tree-based method to directly maximize expected benefit as measured by CATEs across all treatment thresholds.
- We provide an empirical case study to explore the potential for directly maximizing expected benefit compared to optimizing for CATE accuracy. Empirically, across a range of settings, our approach is more sample-efficient and outperforms methods that focus solely on accurate CATE estimation in low-data regimes.

4.2 Problem Set-Up

Setup. We study a setting where the decision maker aims to identify the top $u\%$ of individuals who will benefit most from some resource or treatment, for some value of u that is unknown *a priori*. We assume access to an observational dataset containing n individuals with tuples $S = (\mathbf{x}_i, t_i, y_i)_{i=1}^n$, where each individual i has covariates $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$, assigned treatment $t_i \in \{0, 1\}$, and experiences the observed outcome under the assigned treatment $y_i \in \mathbb{R}$ (for continuous outcomes) or $y_i \in \{0, 1\}$ (for binary outcomes). We follow the potential outcomes framework and define CATE as in **Section 2.2.1**.

Goal. To identify the top $u\%$ of individuals who will benefit (i.e., have the greatest CATE), we seek a function f such that $\forall i, j \in S$ where $\tau_i > \tau_j$, $f(\mathbf{x}_i) > f(\mathbf{x}_j)$. Given

this function, we may then apply a threshold u at inference time to identify the top $u\%$ of individuals for treatment, for any u. Given an ordering of individuals, we evaluate the potential value of it across all thresholds u. Traditional discriminative ranking metrics used to measure ranking in classification, such as the AUROC or concordance index, calculate the proportion of individuals misranked, based on the existence of a pairwise truth function [170, 190]. In our setting, in addition to the pairwise truth function, we also have ground-truth treatment effects. Classification metrics do not take these effects into account and as a result, do not capture the full impact of a misranking on the expected benefit. In our setting, we utilize a metric that incorporates the ground-truth treatment effects, to better understand the expected benefit of a given ranking.

Measuring Expected Benefit. We define how to measure the expected benefit of treating the top $u\%$ of patients in sample S, as identified by model f. Assume that the CATE τ_i is observed and may be used for evaluation. Formally, we define $D_S^u(f)$ as the top $u\%$ of individuals scored by the model, i.e., $D_S^u(f) = \{i | f(\mathbf{x}_i) \geq \psi(\{f(\mathbf{x}_i)_{i \in S}\}, u)\}$, where $\psi(a, u)$ is the uth percentile of the empirical distribution of a. The average benefit from treatment for these individuals is defined as $ATE_S^u(f) = \frac{1}{|D_S^u(f)|} \sum_{i \in D_S^u(f)} \tau_i$. A larger $ATE_S^u(f)$ value corresponds to a function f that better identifies who benefits most from treatment at threshold $u\%$. As in past work, we normalize this value to measure improvement over a random ranking by defining the *targeting operator characteristic (TOC) at* u as the difference between the ATE of the top $u\%$ of patients as ordered by f, and the ATE of treating all individuals, i.e., $TOC_S^u(f) = ATE_S^u(f) - \frac{1}{|S|} \sum_{k=1}^{|S|} \tau_k$ [212]. A value of 0 represents no improvement over random. Finally, to measure this across all treatment thresholds u, we use the *Area Under the Targeting Operator Characteristic* (AUTOC). For an arbitrary function f and a sample S,

$$AUTOC_S(f) = \frac{1}{|S|} \sum_{i=1}^{|S|} TOC_S^{100 * \frac{i}{|S|}}(f)$$

The AUTOC measures the average benefit from treatment of those identified in the top $u\%$ by f, averaged across all thresholds u, relative to the ATE (i.e., the average treatment benefit of a random sample) [212]. Larger values of AUTOC represent more accurate identification of the top $u\%$ of individuals, while an AUTOC of 0 represents a random ranking. The AUTOC may also take negative values if worse than random. While there exist similar metrics, such as the Qini curve, that reweight the objective at different thresholds u, we use the AUTOC due to its strong theoretical properties and unbiasedness when estimated using doubly robust proxies [212].

Causal Identifiability Assumptions. As measuring the AUTOC relies on the true values of τ, it is not identifiable from observational data without additional assumptions.

In line with the majority of work in causal inference, we assume no hidden confounding, overlap, and consistency. These assumptions are sufficient for the identification of causal effects, and hence, are also sufficient for the ranking of causal effects [182, 78, 84]. We discuss the implications of these assumptions at the end of this section.

4.3 Theoretical Analysis

In this section, we study the relationship between accurate CATE estimation and optimal ranking defined by maximizing benefit (**Figure 4.1**). We begin by exploring what it means to maximize benefit across all treatment thresholds as measured by AUTOC. From here, we compare the problem of obtaining accurate CATE estimators to the problem of directly optimizing for AUTOC.

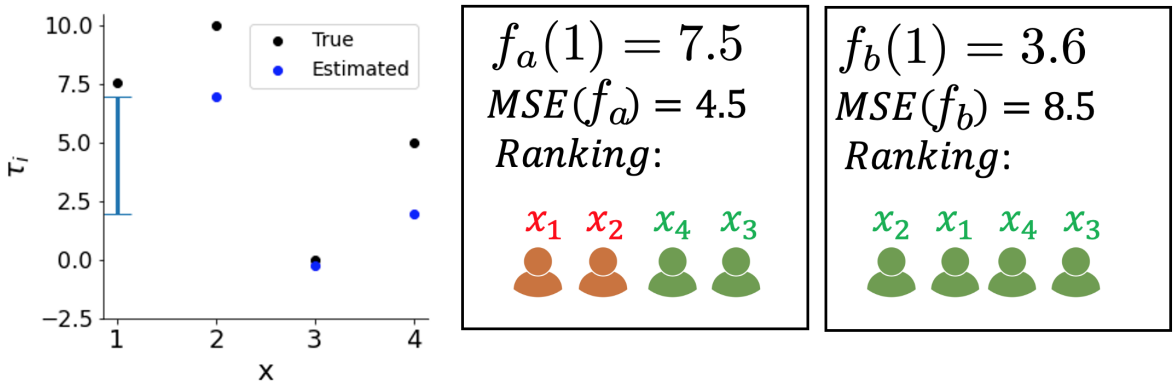

Figure 4.1: A motivating example. Consider four individuals, and a model that has estimated CATEs for individuals $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$. To achieve better mean-squared error (MSE), the model should predict a value close to the true CATE (7.5). However, the model can achieve a perfect ranking by estimating the CATE of the remaining example ($\mathbf{x}_1$) *anywhere* in the interval shown by the blue bar. This illustrates important takeaways from Propositions 1 and 2: 1) we may achieve optimal AUTOC even when the CATE function is not estimated accurately, and 2) a model with better MSE may not result in better AUTOC.

We begin by understanding what it means to maximize AUTOC, where the optimal model is defined as $f^*(\mathbf{x}_i) = \tau_i$ for all $\mathbf{x}_i$.

Claim 1 *Given a function f and a dataset S, $(\forall i, j | \tau_i > \tau_j, f(\boldsymbol{x}_i) > f(\boldsymbol{x}_j)) \leftrightarrow AUTOC_S(f) = AUTOC_S(f^*)$*

Claim 1 states that if a function f correctly orders pairs of examples in terms of their CATE then it will achieve optimal AUTOC performance. Hence, it suffices to find models that are optimal in the ordering of examples to maximize AUTOC. Given this intuition, we

study the relationship between estimating CATEs and AUTOC performance and identify if accurate AUTOC may be easier than accurate CATE estimation.

To do so, we first define $\mathcal{L}_S^M(f) = \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - \tau_i)^2$ as the mean-squared error for CATE estimation for a function f over a sample S. $\mathcal{L}_S^M$ can help measure the performance of a CATE estimation technique as a larger value means worse CATE estimation performance. Next, we introduce the notion of *margins*. We define the margin for point i as $\gamma_i = \min_{j:j\neq i}(f^*(\mathbf{x}_i) - f^*(\mathbf{x}_j))$. The margin measures the extent to which a model can misestimate the CATE without violating an optimal ordering. Given these definitions, we formally study the relationship between CATE estimation accuracy and optimal AUTOC. First, we study the case where a model achieves perfect CATE estimation performance.

Claim 2. *Given a model f and a sample S, $\mathcal{L}_S^M(f) = 0 \rightarrow AUTOC_S(f) = AUTOC_S(f^*)$*

If $\forall i \in S, f(\mathbf{x}_i) = \tau_i$, f is optimal by definition. Hence, a perfect CATE estimator is a sufficient condition for optimal AUTOC. This means that the solution set for optimal AUTOC is at least as large as the solution set for perfect CATE estimation. However, the converse is not true.

Proposition 1. *For a sample S, there exists a function $f \in \mathcal{F}$ such that $AUTOC_S(f) = AUTOC_S(f^*)$, yet $\mathcal{L}_S^M(f) > 0$.*

The proof is simple and can be found in **Appendix B.2**. **Proposition 1** states that a model that achieves perfect AUTOC may obtain arbitrarily poor CATE estimation performance. Hence, accurate CATE estimation is not a *necessary* condition for optimal AUTOC. Our proof technique consisted of creating a function f which is biased in a way that preserves optimal AUTOC but results in a $L^M(f)$ greater than 0. More generally, any function $f \in \mathcal{F}$ that biases each example i by less than half of its margin γ_i will result in optimal AUTOC and non-zero L^M. Hence, the set of solutions that lead to optimal AUTOC may be larger than the optimal solutions for CATE, especially when the ground-truth margin γ between examples is sufficiently large. In these settings, solutions for AUTOC, that simply require a correct ordering of examples, could be easier to learn than the optimal CATE function f^*.

Up to now, we have shown that the set of solutions to AUTOC will be just as large, if not larger, than the set of solutions to CATE accuracy. Optimizing for maximal AUTOC can guide learning towards any of these solutions, potentially resulting in an easier optimization problem. However, our analysis has focused on the sufficiency and necessity of perfect CATEs. We next study the finite sample setting where CATEs may not be estimated perfectly. We show that optimizing for better CATE in these settings does not necessarily lead to better AUTOC performance.

Proposition 2. *For any model f and sample S such that $\mathcal{L}_S^M(f) > 0$, there exists a model g such that $\mathcal{L}_S^M(f) < \mathcal{L}_S^M(g)$ and $AUTOC_S(g) > AUTOC_S(f)$.*

The proof can be found in **Appendix B.2**. Importantly, **Proposition 2** says that a better CATE estimator may not result in a greater AUTOC. **Hence, optimizing for CATE accuracy does not necessarily translate to better AUTOC in settings where the CATE function cannot be estimated well.**

Given that the solution set for optimal AUTOC is larger than the solution set for perfect CATE estimation, and better CATE accuracy does not necessarily translate to better AUTOC, we hypothesize that optimizing directly for AUTOC, at the cost of CATE estimation performance, could lead to better performance as measured by ranking for maximal benefit. We expect this will hold in low and finite sample settings, where estimating the CATE function exactly might be challenging.

In this work, we stop short of characterizing the complexity of the problem of optimizing for AUTOC compared to optimizing for accurate CATEs. Past work has shown that the problem of CATE estimation scales with the complexity of the true underlying CATE function [3]. We hypothesize that the problem of optimizing for AUTOC does not scale with the complexity of the CATE estimation function, as we have shown that accurate CATEs are not a necessary condition for accurate AUTOC. However, understanding what the complexity of optimizing for AUTOC depends on remains difficult. A majority of work in understanding the complexity of ranking simply considers the ordering of different examples and does not consider how differences in outcomes may factor into the objective [170]. For example, recent work in uplift modeling, under the assumption of data from a randomized controlled trial and binary outcomes, showed that the complexity of their objective function could be decomposed into multiple AUROC bounds by viewing their objective as a bipartite ranking problem [21]. In our setting with observational data and potentially continuous outcomes, where the magnitude of CATEs factors into the objective function, their results do not apply.

One challenge in studying the complexity of the AUTOC is the non-differentiability of the objective function. An interesting future direction could be to study the properties of a surrogate of the AUTOC, an approach considered in the related field of learning to rank when understanding the complexity of a different objective function [194]. However, in the remainder of this section, we instead test our hypothesis that optimizing directly for AUTOC, at the cost of CATE estimation performance, could lead to better ranking performance. To do so, we seek approaches that optimize for AUTOC directly in the next section

4.4 Methods

Up to now, we have shown that the solution set for optimal AUTOC is at least as large as the solution set for accurate CATEs, and may be larger. Moreover, in finite settings, a

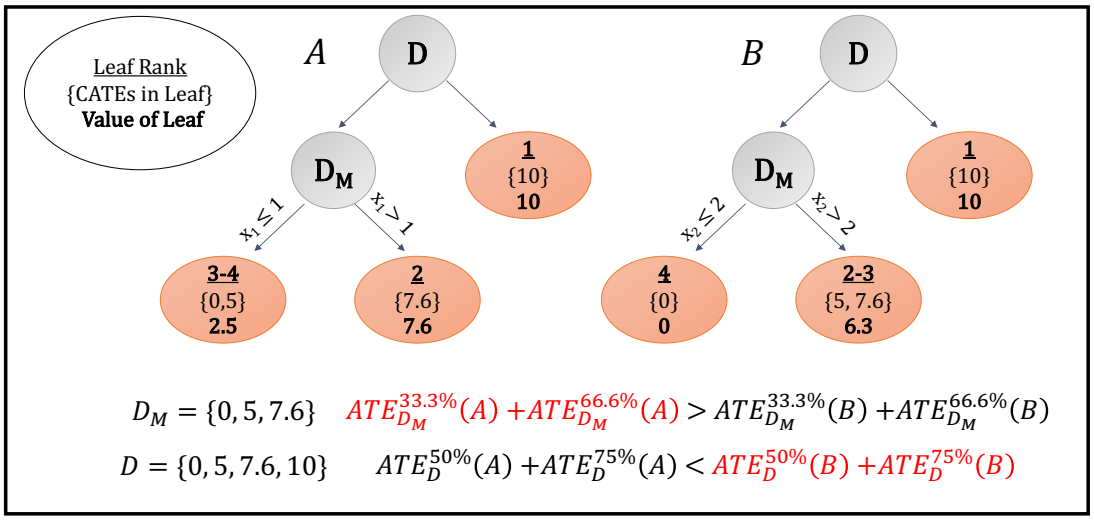

Figure 4.2: The importance of global splits. We define a subtree with a group of data that have CATEs D, in which we aim to split at decision node M, resulting in either tree A or B. A 'local' split based on only data with CATEs D_M results in tree A, as the sum of ATE^u at the first two thresholds $(7.6 + \frac{7.6+2.5}{2})$ is greater than that of tree B $(6.3 + 6.3)$, with the ATE^u at all other thresholds being equal. Globally, tree B is optimal as the sum of ATE^u for the second and third threshold $(\frac{10+6.3}{2} + \frac{10+5+7.6}{3})$ is greater than that of tree A $(\frac{10+7.6}{2} + \frac{10+7.6+2.5}{3})$. Many small differences can result in drastically different performance, so it is important to consider the entire decision tree when selecting splits.

perfect CATE estimator may not directly translate to a better AUTOC. We hypothesize that in some settings, such as low sample settings, optimizing directly for AUTOC may result in better treatment allocation. To test this hypothesis, we next develop a technique for explicitly optimizing for AUTOC across a sample S.

Optimizing For and Calculating AUTOC. While we aim to maximize AUTOC for a sample S, this is made difficult due to the non-differentiability of the AUTOC. Thus, we propose a tree-based approach. Tree-based techniques can be used to tackle arbitrary optimization problems through the use of novel splitting rules. A splitting rule for creating new nodes in a decision tree is not required to be differentiable. We utilize decision trees to directly optimize for AUTOC over a sample S. Moreover, we extend splitting rules to use training examples beyond those seen in the current node in the tree, inspired by past work in learning to rank [82].

To begin, for any decision tree T, the AUTOC for a sample S can be calculated as follows:

1. Assign a score $T(\mathbf{x}_i)$ for each individual i in S based on the average outcome in the leaf node that the example $\mathbf{x}_i$ is partitioned into.

2. Calculate $AUTOC_S(T)$ using the scores $T(\mathbf{x}_i)$ as model outputs. To handle ties where multiple examples have the same predicted score, average across all possible orderings to simulate breaking ties at random.

Building Decision Trees to Maximize AUTOC. We propose an approach for building a tree T to optimize for AUTOC. We first assume we have access to τ_i for all individuals in our sample S, later relaxing this assumption. At any decision node M in a tree, we denote the current samples at that node as S_M and the current tree as T^M. Denote $T^M_{k,v}$ as the tree when the current decision node M is split into two leaf nodes based on the feature k and value v. Traditional regression trees select the best splitting k and v that splits the data into $S_{M_1^{k,v}}$ and $S_{M_2^{k,v}}$ by minimizing the weighted variance of the outcomes over resulting nodes. We propose finding k, v by maximizing the AUTOC for the full sample S. More formally, at each split, we solve the following optimization problem: $k^*, v^* = \arg\max_{k,v} AUTOC_S(T^M_{k,v})$. We use the current estimates at the leaf nodes throughout the decision tree (i.e., the average τ_i value of the leaf node that each example is currently placed at) to calculate the AUTOC. In utilizing these *'global' splits*, we overcome the limitations of local splits (**Figure 4.2**). While all data is considered at each split, the tree is still grown greedily, only slightly increasing computation time (i.e., this is *not* a globally optimal decision tree). The order in which the 'global split' tree is built is important, as the values of all nodes are used at each split. We build decision trees in a breadth-first manner to ensure every portion of the tree is

growing equally, and splits at each node are made using nodes at similar depths. Overall, by utilizing a splitting rule to maximize the AUTOC, we optimize for our end goal of learning accurate rankings for treatment allocation based on maximizing expected benefit. Given this training procedure, we bootstrap our data multiple times and build many decision trees to overcome overfitting and improve performance, as in traditional random forests [26]. At inference time, each test sample is evaluated by each tree, and the outputs are averaged. These estimates are used to rank test data.

Using Doubly Robust Proxies for Training. Relaxing the assumption of oracle access to the ground truth CATE τ_i in our training sample, we use a *doubly robust proxy* of the treatment effect $\tilde{\tau}_i$ for each individual i. The doubly robust estimate is defined as $\tilde{\tau}_i = \hat{m}(\mathbf{x}_i, 0) - \hat{m}(\mathbf{x}_i, 1) + \frac{t_i - \hat{e}(\mathbf{x}_i)}{\hat{e}(\mathbf{x}_i)(1-\hat{e}(\mathbf{x}_i))}(y_i - \hat{m}(\mathbf{x}_i, t_i))$, where $\hat{e}(\mathbf{x}_i)$ is an estimate of the propensity score conditioned on observed covariates, and $\hat{m}(\mathbf{x}_i, t_i)$ is an estimate of the expected outcome given an individual's covariates and treatment assignment [37, 105]. The nuisance parameters represent nonparametric estimates of the ground-truth propensity score and potential outcome functions. Under our assumptions, $E[\tilde{\tau}_i|\mathbf{x}_i] \to \tau_i$ as $n \to \infty$. To calculate the AUTOC, we first calculate the ATE at each threshold using these proxies in place of the true CATEs, i.e., $\widetilde{ATE}^u_S(T) = \frac{1}{|D^u_S(T)|}\sum_{i \in D^u_S(T)} \tilde{\tau}_i$. From here, we calculate the TOC and the AUTOC respectively as $\widetilde{TOC}^u_S(T) = \widetilde{ATE}^u_S(T) - \sum_{k=1}^{S} \tilde{\tau}_k$ and $\widetilde{AUTOC}_S(T) = \frac{1}{|S|}\sum_{i=1}^{|S|} \widetilde{TOC}_S^{100*\frac{i}{|S|}}(T)$. Importantly, $\widetilde{AUTOC}_S(T)$ calculated using $\tilde{\tau}_i$ in place of the true τ_i is an asymptotically unbiased and normal estimate of the true $AUTOC_S(T)$ under mild conditions [212]. In a first stage, these DR proxies can be built using cross-fitting. Then, when making a split at decision node M, we find the k, v pair that maximizes $\widetilde{AUTOC}_S(T^M_{k,v})$. model that directly maximizes the AUTOC, as proposed in the previous section, using the doubly robust proxy will also, in expectation, maximize the true AUTOC.

4.5 Experiments & Results

Empirically, we test our hypothesis that directly optimizing for AUTOC can outperform models focused on CATE estimation in low-sample sample settings. First, we describe our experimental setup and baseline methods. From here, we present the datasets used in our experiments, as well as the evaluation metrics used to measure performance. We then present results comparing the techniques across both datasets to understand the efficacy of the proposed methodology.

4.5.1 Experimental Set-Up

Baseline. As a baseline, we compare to a strong CATE estimation baseline from past work known as the DR-Learner [105]. The doubly robust proxy $\tilde{\tau}_i$ for each example can only be built for individuals for whom treatments and outcomes are observed. Hence, on a new set of examples for whom the treatment and outcome are not observed, these proxies are not available. To overcome this, the DR-Learner learns a mapping from an example's covariates to an estimate of the CATE by regressing $\tilde{\tau}_i$ on an individual's covariates. Formally, the DR-Learner is a two-stage approach similar to our proposed technique. However, in the second stage, the model is trained to accurately estimate the doubly robust proxy using traditional metrics such as mean-squared error. To build the DR-Learner, we train a random forest algorithm similarly to our proposed method. However, at each decision node M, k, v are selected to minimize the balanced variance of outcomes $\tilde{\tau}_i$; the split at decision node M with data-points S_M can be defined as $argmin_{k,v} \frac{|S_{M_1^{k,v}}|}{|S_M|} Var(\{\tilde{\tau}_i\}_{i \in S_{M_1^{k,v}}}) + \frac{|S_{M_2^{k,v}}|}{|S^M|} Var(\{\tilde{\tau}_i\}_{i \in S_{M_2^{k,v}}})$. At inference, outputs in each tree are aggregated by taking the average doubly robust outcome. Although numerous other CATE estimation models have been proposed recently, we opt for a strong baseline approach that is similar to our proposed method to test our primary hypothesis. We use the same doubly robust proxies for training for both methods such that any observed differences between the two approaches can be attributed to differences in the splitting criteria. We tune the number of trees, the proportion of data in each tree, the maximum depth of each tree, the threshold for improvement, the minimum number of samples needed for a split, and the minimum number of samples at a leaf as hyperparameters for both models (see **Appendix B.4** for more detail).

Datasets. While CATE estimation arises frequently in practice, validating these techniques in real data requires close collaboration with domain experts since there is no well-accepted approach to evaluating these models without ground truth. Hence, as a first step, we focus on existing synthetic datasets in which the counterfactual is available. We test our proposed approach using synthetic data-generating procedures adapted from past work [14, 33]. Specifically, we generate two datasets:

Dataset 1	**Dataset 2**
$\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_{10x10})$,	$\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_{10x10})$
$t_i \vert \mathbf{x}_i \sim Bern(\frac{1}{1+e^{-x_{i,3}}})$,	$t_i \vert \mathbf{x}_i \sim Bern(\frac{1}{1+e^{-x_{i,3}}})$,
$\epsilon_i \vert \mathbf{x}_i, t_i \sim \mathcal{N}(0,1)$,	$\epsilon_i \vert \mathbf{x}_i, t_i \sim \mathcal{N}(0,1)$,
$\tau_i \vert \mathbf{x}_i = ((x_{i,1})_+ + (x_{i,2})_+ - 1)/2$,	$\tau(\mathbf{x}_i) = 1 + 2\vert\mathbf{x}_{i,4}\vert + \mathbf{x}_{i,10}^2$,
$y_i \vert \mathbf{x}_i, \tau_i, \epsilon_i, t_i = \max(0, x_{i,3} + x_{i,4}) + t_i\tau_i + \epsilon_i$	$y_i \vert \mathbf{x}_i, \tau_i, \epsilon_i, t_i = 5(2 + 0.5\sin(\pi\mathbf{x}_{i,1})$ $- 0.5\mathbf{x}_{i,2} + 0.75\mathbf{x}_{i,3}\mathbf{x}_{i,9}) + t_i\tau_i + \epsilon_i$

In **Dataset 1**, the ground truth τ_i function is built by thresholding certain covariates in each individual. In doing so, we can create different groups of individuals with different treatment effects, resulting in large margins between individuals. This is a setting in which we expect our proposed approach to perform well. Using **Dataset 2**, we test our approach in a more complex setting in which the underlying CATE and outcome functions involve more non-linear terms.

Evaluation Metrics. We assess the performance of our proposed approach and the baseline on both datasets, each with 30 unique replications for training and testing. To understand how the proposed method performs with varying amounts of training data, we sweep the amount of training data N through $\{100, 250, 500, 1000\}$, while keeping the test set size fixed at 5000. We focus on a low-sample regime as in many domains, obtaining interventional trial data is challenging. For example, in the field of healthcare, many diseases are rare and many patient populations are less represented in the data. Due to this, many problems in the field of healthcare are plagued with issues due to a limited number of examples [45, 35]. Efficiently learning accurate rankings in these regimes remains imperative. We evaluate the performance of the methods on held-out test sets in terms of the AUTOC, reporting the median and interquartile range (IQR) across all 30 replications. Additionally, since each dataset may have different optimal AUTOC values, we report the number of times the proposed method outperforms the baseline across the 30 random seeds. We also evaluate the ATE^u, which helps in understanding the difference in realized benefit at specific thresholds. We test $u \in \{10, 20, 30, 40, 50\}$, to evaluate realistic settings in which the treatment can only be administered in a fraction of individuals. Relative to the baseline, we report the median improvement in ATEs at each threshold across 30 replications. For completeness, we report both the % of replications the proposed method outperforms the baseline across the 30 random seeds for each u and TOC^u performance across all thresholds in **Appendix B.5**.

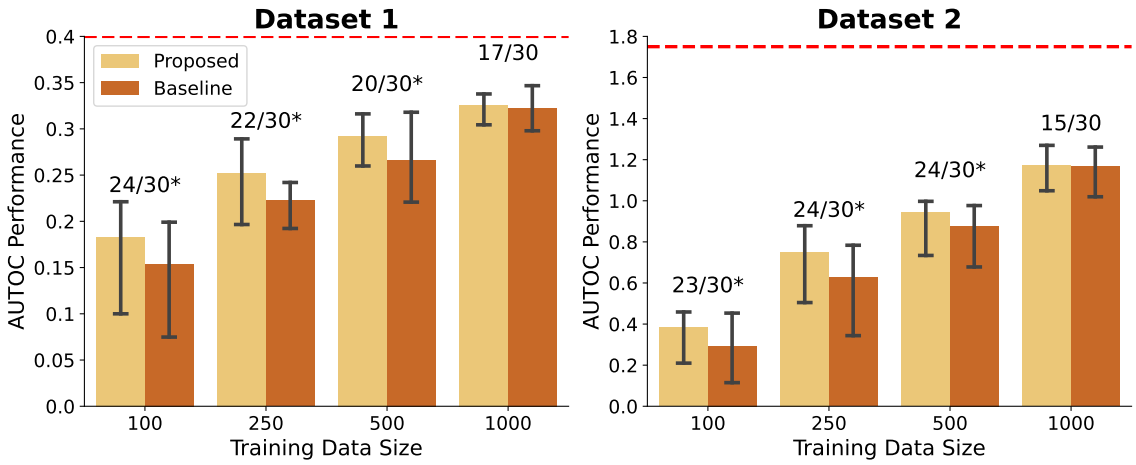

Figure 4.3: Median and IQR AUTOC as well as how many times the proposed method outperforms the baseline across 30 replications. Asterisks represent scenarios where the proposed method significantly outperforms the baseline technique as measured with the Wilcoxon signed rank test ($\alpha = 0.05$). The maximum AUTOC achievable is indicated by the red dashed line. At low sample sizes, the proposed method outperforms the baseline.

4.5.2 Results

AUTOC Performance: On both datasets, at low-sample settings, our proposed approach outperforms the baseline CATE estimation technique on a large majority of replications (N = 100: 24 and 23 /30 replications, N = 250: 22 and 23/30 replications, respectively) (**Figure 4.3**). As the sample size increases, both approaches begin to perform similarly. In data-rich settings (N = 1000), the baseline may be preferable due to its simplicity. Notably, this trend holds even when using local splits (**Appendix B.5**). Empirically, local splitting results in similar splits early on in the tree-building process, but diverges at greater depths. More recently, researchers have proposed an honest framework for training decision trees for CATE estimation [13]. In the honest framework, when training, only half of the data is used to create the splits, and the other half is used to impute outcomes at each leaf node during inference. To show that our approach is robust to the honest framework, we repeat our analysis and show that our model still outperforms the baseline technique in a low-sample setting (**Appendix B.5**). For completeness, we also compare our approach to that of [223] in **Appendix B.5** and show that our proposed method for directly optimizing for AUTOC outperforms this baseline significantly.

ATE^u Performance: Evaluating the value of a learned ranking at specific treatment thresholds (i.e., ATE^u), our proposed technique outperforms the baseline in low-data settings when treating between 10% and 50% of individuals (**Figure 4.4**). Across training set sizes

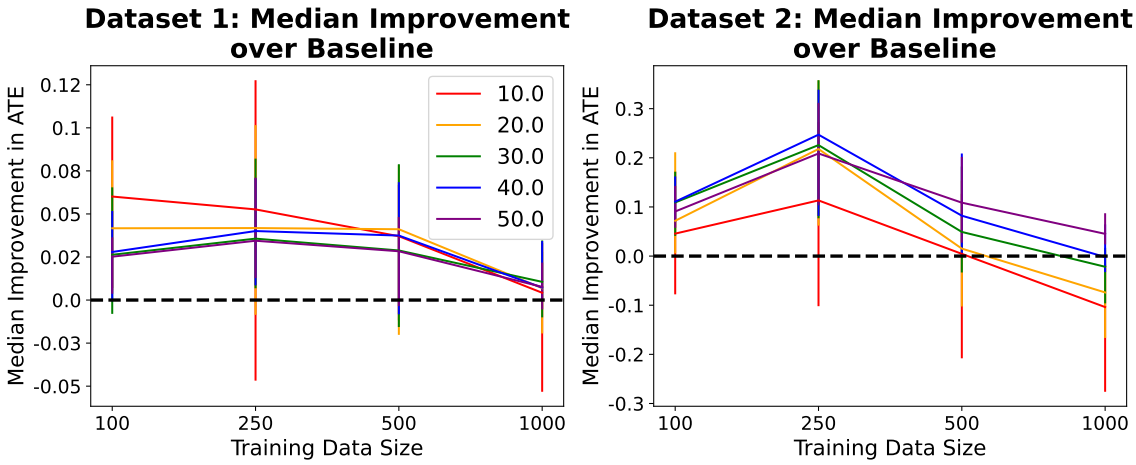

Figure 4.4: The median and IQR of improvement of the proposed approach over the baseline in ATE^u across potential thresholds u. Our model excels across treatment thresholds at low-data settings, despite not being trained for a particular treatment threshold. With more training data ($N = 1000$), the efficacy of our model is shown at higher treatment thresholds.

of N = 100 to N = 500, the proposed training scheme consistently outperforms the baseline in **Dataset 1**, with improvements in ATEs of up to 0.06. Our model continues to perform well across thresholds in **Dataset 2**, outperforming the baseline at almost all thresholds in low-data settings, with median ATE improvements of up to 0.25. With more training data (N = 1000), the baseline slightly outperforms the proposed technique at lower treatment thresholds, but the proposed approach demonstrates efficacy at higher treatment thresholds. At higher sample sizes, the worse performance at lower treatment thresholds balances out with the better performance at higher thresholds, resulting in similar overall AUTOCs. In addition, our proposed method outperforms the baseline technique in terms of ATE^u in up to 80% of replications and outperforms the baseline at thresholds beyond $u = 50$ consistently as well (**Appendix B.5**).

Contextualizing Results: To understand the potential impact of our direct optimization of ranking, we introduce an evaluation that emulates a setting where treatment improves the probability of survival. We shift and normalize CATEs and outcomes in both datasets such that the maximum values are 1 and 0. An outcome of 1 represents a 100% chance of survival and an outcome of 0 represents a 0% chance of survival, and a τ_i of 1 means that treatment completely reduces the likelihood of death, whereas a τ_i of 0 means that treatment does not affect survival. The expected lives saved at any threshold u can then be calculated as the ATE for individuals allocated the treatment, as this is exactly the expected improvement in mortality in those treated. We then normalize these values by the maximum possible lives saved at u given a perfect ranking, which we denote as *% lives saved at u*. We

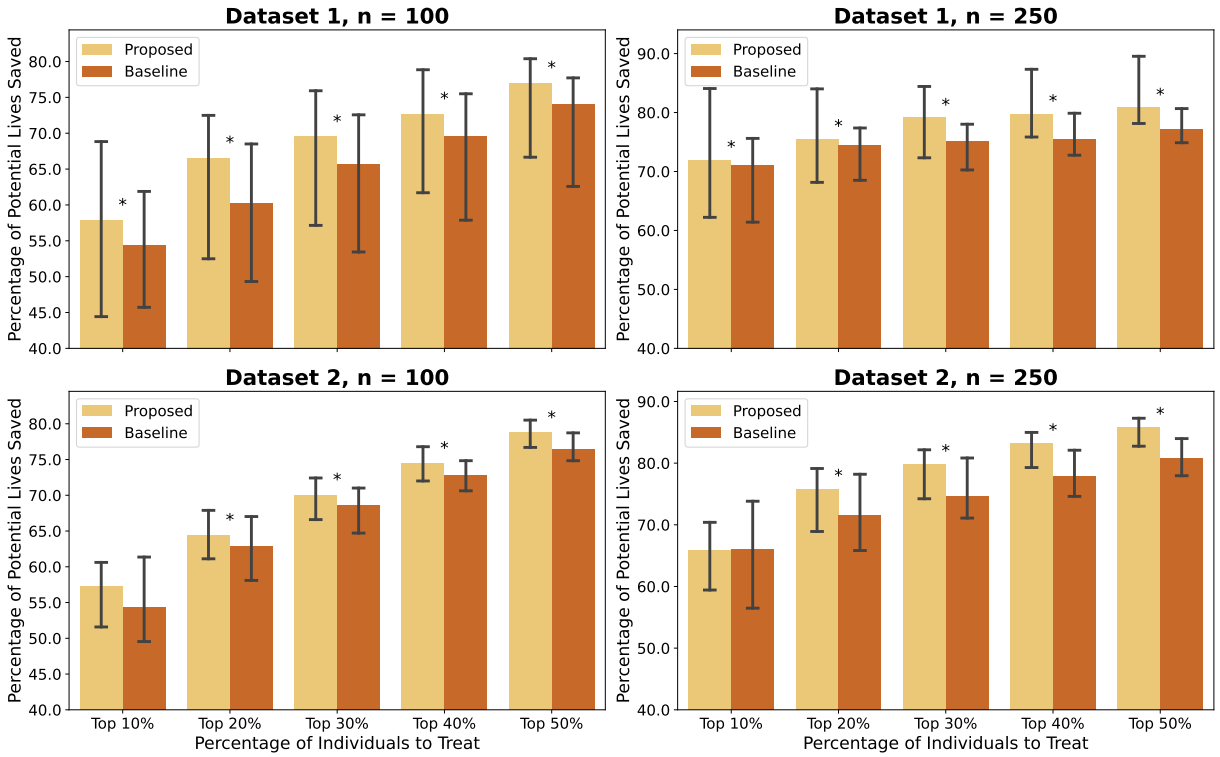

Figure 4.5: Median and IQR of the percentage of potential lives saved compared to the oracle across different thresholds in low data settings for **Dataset 1** (top) and **Dataset 2** (bottom). Asterisks represent scenarios in which the proposed method significantly outperforms the baseline technique as measured using a Wilcoxon signed rank test ($\alpha = .05$). The proposed method consistently outperforms the baseline technique in terms of lives saved, with up to a 6.4% increase.

perform this analysis across all training data settings and thresholds $u \in \{10, 20, 30, 40, 50\}$.

Across both datasets, the proposed method consistently outperforms the baseline technique in terms of % lives saved (**Figure 4.5**). At $u = 30$, the proposed method consistently outperforms the baseline (**Dataset 1**: N = 100: 69.5% vs. 65.6% and N = 250: 79.1% vs 75.2%, **Dataset 2**: N = 100: 70.0% vs. 68.6% and N = 250: 79.8% vs 74.6%). In data-rich settings, the proposed method matches the performance of the baseline or performs only slightly worse (**Appendix B.5**). Overall, this evaluation demonstrates the potential that the proposed method could have in resource-constrained settings.

4.6 Conclusion and Discussion

In this chapter, we study the problem of intervention allocation. Past work often considers solving this problem by accurately estimating CATEs from observational data to help triage

individuals. However, in situations where all one needs is a ranking of who is more likely to benefit, there exists an objective mismatch between what one is optimizing for and what one needs. Our work builds on past research focused on the disconnect between exact causal effect estimation and the ultimate goal of augmenting downstream decision-making [52, 14, 96]. We show that optimizing for CATE accuracy, while sufficient, is not necessary for optimal expected benefit, and that the set of solutions for accurate ranking is just as large, if not larger, than the set of solutions for accurate CATE estimation. We also show that models achieving better CATE performance may not always translate to better ranking. Based on this analysis, we hypothesize that optimizing directly for ranking can outperform methods focused on minimizing mean squared error. To test this hypothesis empirically, we propose an approach for optimizing ranking in this context and test our hypothesis empirically. With respect to triaging individuals to maximize benefit, our proposed approach achieves strong empirical performance and better sample efficiency compared to a baseline CATE estimation method across two synthetic datasets.

Our study is not without limitations. First, due to the inability to observe ground-truth CATEs in real observational data, we could not explore performance on real datasets. While results on different synthetic datasets help demonstrate the initial efficacy of the proposed method and problem setting, future work should consider how to effectively validate these models in a multitude of real settings. In particular, it remains important to carefully validate these algorithms in close collaboration with domain experts before they are used to inform decision-making. Second, as our work focuses on the problem of resource allocation under constraints, we consider a utilitarian solution to the problem of resource allocation, such that we maximize the expected benefit across all treatment thresholds. However, decisions on resource allocation are often multi-faceted and require considerations beyond simply maximizing the expected benefit for the full population [195, 151]. For example, there exist many ethical constraints that may be considered when allocating sparse interventions, as recently shown during the COVID-19 pandemic [216]. Our work is intended to study one tool that may be used to augment this decision-making, which may also be combined with other important societal considerations. In addition, like most work in causal effect estimation, we make three common assumptions to ensure the identifiability of CATEs: 1) unconfoundedness, 2) consistency, and 3) overlap. These assumptions ensured that our doubly robust proxy was identifiable and could be used for training. However, as the problem of accurate resource allocation based on benefit does not require the ground-truth CATEs to be estimated perfectly, there exists a potential to relax these assumptions and learn how to optimize for accurate rankings [51]. We further discuss how these assumptions may be relaxed at the end of this dissertation. Finally, our proposed approach relies on a proxy for

learning. Future work could consider how to directly optimize for AUTOC that overcomes the need for a proxy on the training set. However, our approach still shows the empirical benefits of directly optimizing for AUTOC in the downstream estimator, as both our proposed approach and the baseline rely on the same proxy during training.

Despite the obvious relationship to triage, to the best of our knowledge, we are the first to consider the efficacy of directly optimizing for maximum benefit in treatment allocation under variable resource constraints in observational data. Overall, our work represents an important step for bridging the theory and practice of resource allocation techniques.

CHAPTER 5

Challenging Implicit Assumptions of Theory Through Empirical Evidence in CATE Estimation

5.1 Introduction

In precision medicine, there exist many situations in which a ranking of individuals by benefit may not be sufficient. For example, when deciding on a treatment rule for a particular individual, accurately estimating CATEs can be critical to weigh the benefit of treating against other potential alternatives. Hence, estimating CATEs has an immense potential to improve different aspects of clinical decision-making [120]. To address the challenge of confounding, a number of different learning algorithms have been proposed that can broadly be grouped into three categories [145, 214]. The first category of techniques focuses on only using a model of the outcome during training by using relevant confounders as input covariates (such as plug-in estimators or g-computation). However, these methods may be inefficient in high-dimensional settings when the level of confounding is large [165, 30, 179, 22]. Researchers hence often consider a second category of models that incorporate estimates of the propensity score, the probability of receiving the treatment, during training. Propensity scores provide a single scalar value that can balance treatment groups, often making them preferred over simpler outcome-based techniques [165, 179]. However, such methods may fail when the propensity score cannot be estimated accurately. To overcome this, recent work has considered a third category of approaches that adjust for both the propensity score and an estimate of the potential outcome during training [142, 105, 43]. By using both, these techniques can derive theoretical guarantees for accurate CATE estimation regardless of errors in one or both of the estimates [142, 105, 56]. For example, doubly robust approaches are robust to misspecification of either the outcome model or the propensity score model. Due to this, these adjustment techniques have become increasingly favored over other

approaches [142, 105, 59].

While theory supports the choice of techniques that incorporate both estimates of the outcome and the treatment assignment, this theory often does not consider finite sample performance, resulting in a mismatch between theory and practice. Despite the multitude of CATE estimation techniques, there is little comprehensive empirical evidence to guide practice, especially in the modern context of deep learning. Direct comparisons between different learning algorithms used in the literature are made even more difficult as base learners used to train these algorithms can vary from linear regression to more complex neural network approaches. To date, empirical investigations of CATE estimation techniques have been limited to only a narrow set of CATE estimation approaches, do not focus on a single strong base learner, or do not explore performance across a wide range of relevant settings, including different levels of confounding and noise in the estimates of the propensity score [43, 145, 109, 138].

Leveraging both synthetic and benchmark semi-synthetic datasets, in this chapter, we explore the performance of popular CATE estimation techniques within each category across a wide range of settings, including different levels of confounding and propensity score errors. To provide a fair comparison across techniques while considering the modern context of deep learning, we investigate these approaches using an increasingly popular neural-network base learner [182, 220, 219, 75, 12, 38]. In contrast to some past work, this also allows us to compare to popular CATE estimation techniques that are specifically built using neural networks [48]. Our empirical analyses highlight the failure modes of many popular CATE estimation techniques. Overall, we find that many popular CATE estimation approaches fail to consistently outperform simpler approaches using only a model of the outcome during training, even when given access to ground-truth propensity scores. Furthermore, our empirical analyses highlight the sensitivity of many techniques, including doubly robust techniques, to errors in the propensity score. Our findings offer valuable insights and important considerations for researchers using CATE estimation techniques across many applications, including healthcare.

5.2 Background and Related Work

We continue with the problem set-up and assumptions described in **Section 2.2**. We focus on evaluating the most popular CATE estimation techniques. All techniques utilize observed confounders $\mathbf{x}_i$ as input to a machine learning model, with the goal of learning accurate CATEs. Indirect methods learn CATE estimates through learning models $\hat{f}^1, \hat{f}^0$ that map $\mathbf{x}_i \rightarrow Y_i(1)$ and $Y_i(0)$ respectively, and estimate the CATE as the difference

between the estimated potential outcomes. Assuming continuous outcomes, these methods are trained to minimize loss on the observed outcomes using the following objective: $\mathcal{L}_o = \frac{1}{n}\sum_{i=1}^{n} w_i(y_i - \hat{f}^{t_i}(\mathbf{x}_i))^2$, where w_i are weights that may be specific to a particular algorithm. Direct learners aim to estimate a model f that maps the covariates to the CATE directly, i.e., $\mathbf{x}_i \rightarrow \tau_i$. These methods first learn relevant nuisance parameters, such as the propensity score or the observed outcomes, in a first stage to build a proxy estimate $\tilde{\tau}_i$ of the CATE in the training set. In a second stage, these methods learn a CATE estimator by building a model $\hat{f}$ which maps input covariates to the proxy from the first stage by minimizing $\mathcal{L} = \frac{1}{n}\sum_{i}^{n} w_i(\tilde{\tau}_i - \hat{f}(\mathbf{x}_i))^2$, where weights may be specific to a particular algorithm. Across both direct and indirect methods, techniques differ primarily in how they adjust for confounding during training. Popular methods can be broken down into three main categories: outcome-based models, propensity score adjustment models, and models that adjust using both the propensity score and outcome estimates. We next consider different implementations of CATE estimators across these three approaches (**Table 5.1**). For all methods, we consider using neural networks as a base learner to optimize the model.

Table 5.1: Overview of all methods considered.

Model	Adjustment	Learning Type
TARNet [182]	Outcome-Based	Indirect
X-Learner [112]	Outcome-Based	Direct
Weighting	Propensity Score	Indirect
Matching	Propensity Score	Indirect
R-Learner [142]	Propensity Score + Outcome Estimate	Direct
DR-Learner [105]	Propensity Score + Outcome Estimate	Direct
DragonNet [185]	Propensity Score + Outcome Estimate	Indirect

1. Outcome-Based Models (or Plug-In Methods). As discussed in **Section 2.2**, two early techniques under this category were the S-Learner and T-Learner. We continue with describing new techniques proposed in recent years that build upon these early techniques. We focus on two popular approaches.

TARNet improves on the S and T-Learner through a multi-task framework [182, 42]. TARNet is an indirect algorithm that learns accurate estimates of the observed outcome by minimizing $\mathcal{L}_o$ using a multi-task neural network architecture with a shared representation Φ and two separate outcome network heads h^1, h^0 for each potential outcome. The potential

outcomes are thus estimated as $\hat{f}^t = h^t(\Phi(\mathbf{x}_i))$ for $t \in \{0, 1\}$. The strength of this architecture is supported by generalization bounds for CATE estimation, in which the learning bound for CATE is determined by the more complex potential outcome function [3]. We set w_i to provide equal supervision to both outcome heads as in past work [182].

X-Learner is a direct CATE estimation algorithm that simply relies on estimates of the potential outcomes [112]. In the first stage, estimates of the potential outcome $\hat{\mu}_{1i}, \hat{\mu}_{0i}$ are obtained by learning f^1 and f^0. Given these estimates, two new datasets D_1, D_0 are formed for the treatment and control groups, with modified outcomes $y_i - \hat{\mu}_{0i}$ and $\hat{\mu}_{1i} - y_i$ serving as proxy CATEs for treatment and control individuals separately. In the second stage, these datasets can then be used to train two direct CATE estimators from D_1, D_0 respectively by optimizing for $\mathcal{L}$ with the proxy CATEs. Finally, at inference, these CATE estimators can be combined for a final estimate using a weighted average. The X-Learner uses information from each treatment group to derive better estimators for the other cohort. To encourage this further, rather than training two separate models using D_1 and D_0, we train a multi-task neural network architecture similar to TARNet such that each head corresponds to the CATE estimate learned from the treatment and control groups separately, but the learned representation is shared and learned from all training examples. In **Appendix C.2**, we show that this modification for the X-Learner substantially improves performance over the traditional X-Learner. As suggested by the original authors, we use estimates of the propensity score to weigh the two CATE estimators. Though the X-Learner uses the propensity scores during inference, it does not adjust for it during training like the techniques in the next section.

2. Adjustment Using Only Propensity Scores During Training. To implement these techniques, we consider indirect estimators that build off TARNet and use propensity scores estimated in a first stage. We focus on two common approaches for propensity score adjustment. In **Appendix C.4**, we also describe a direct approach that inversely weights the observed outcome using the propensity score to create a proxy for CATE. However, we omit this method from this section due to its poor empirical performance.

Weighting reweights the loss function of the observed outcomes $\mathcal{L}_o$ using the inverse of the propensity score [12, 224]. Specifically, we reweight the loss function by using stabilized weights, which multiply the traditional weights by the marginal probability of the observed treatment and reduce variance [16, 78]. Formally, we reweight the loss function using weights $w_i = P(t = t_i)(\frac{t_i}{\hat{e}_i} + \frac{1-t_i}{1-\hat{e}_i})$.

Matching creates a matched set M_i for each individual. For a particular individual i, we consider including all individuals within a specific distance c from i in terms of the propensity score in the opposite treatment group, where c is defined as 0.2 times the standard deviation

of the logit of the propensity score in the population, due to strong theoretical guarantees with this cut-off [166, 15]. We impute unobserved outcomes $\hat{y}_i^{1-t_i}$ for individual i by taking the weighted average of outcomes for individuals in M_i, or $\hat{y}_i^{1-t_i} = \frac{\sum_{j:M_i} w_j y_j}{\sum_{j:M_i} w_j}, w_j = \frac{1}{|\hat{e}_i - \hat{e}_j|}$. w_j weights examples based on their propensity score distance, allowing individuals closer in propensity score to contribute more to the estimated outcome and removing training examples without any relevant nearest neighbor from training. Given these imputed outcomes, the model is trained to accurately estimate both the observed and estimated unobserved outcome using a composite loss of $\mathcal{L}_o + \alpha \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i^{1-t_i} - \Phi(h^{1-t_i}(\mathbf{x}_i))^2$, where $\alpha > 0$.

3. Adjustment Using Propensity Score and Outcome Estimates During Training. We consider three approaches in this category. The first two approaches are direct approaches, while the third approach is an indirect approach. In **Appendix C.4**, we also consider the U-Learner. However, we omit this method from this section due to its poor empirical performance.

R-Learner uses both an estimate of the propensity score e_i as well as an estimate of the conditional outcome $m_i = E[y_i|\mathbf{x}_i]$ learned in a first stage to create a proxy outcome $\tilde{\tau}_i = \frac{y_i - \hat{m}_i}{t_i - \hat{e}_i}$, where $E[\tilde{\tau}_i|\mathbf{x}_i] = \tau_i$ [142, 112]. The proxy is regressed on the covariates by minimizing $\mathcal{L}$, with weights set to $(t_i - \hat{e}_i)^2$ [142, 145, 55]. We utilize a feed-forward neural network in the second stage to learn the CATE estimator. The R-Learner can achieve similar error bounds to an oracle approach that has ground-truth estimates of the propensity score and conditional outcome, regardless of the true accuracy of these estimates. This property makes it especially attractive for practical use when building CATE estimators.

DR-Learner is a direct approach that uses estimates of both the potential outcome functions μ_{1i}, μ_{0i} obtained from $\hat{f}^1, \hat{f}^0$ as well as the propensity score e_i obtained from a first stage [105]. The second-stage proxy is defined as $\tilde{\tau}_i = \frac{t_i - \hat{e}_i}{\hat{e}_i(1-\hat{e}_i)}(y_i - \hat{\mu}_{t_i i}) + \hat{\mu}_{1i} - \hat{\mu}_{0i}$, also known as the augmented inverse propensity weighting (AIPW) proxy. A CATE estimator is then learned by minimizing $\mathcal{L}$ using this proxy. The DR-Learner is doubly robust, in that only the propensity score estimator or the potential outcome estimator needs to be specified correctly for asymptotically unbiased CATEs, making it a popular choice for CATE estimation. We utilize a feed-forward neural network in the second stage to learn the CATE estimator.

DragonNet is a single-stage learner which learns and adjusts for the potential outcomes and propensity scores during training, rather than in separate stages [185]. DragonNet modifies the TARNet architecture by adding a simple linear map π from the learned shared representation Φ that is trained to accurately estimate the propensity score in combination with $\mathcal{L}_o$. In doing so, the model is encouraged to learn a representation that is predictive of the treatment assignment. Moreover, inspired by the field of targeted regularization, an extra model parameter ϵ is introduced along with a regularization term defined as $\frac{1}{n} \sum_{i=1}^{n} (y_i - Q)^2$,

where Q is a perturbed version of the estimated outcome defined as $h^{t_i}(\Phi(\mathbf{x}_i)) + \epsilon[\frac{t_i}{\pi(\Phi(\mathbf{x}_i))} - \frac{1-t_i}{1-\pi(\Phi(\mathbf{x}_i))}]$. Minimizing this regularization term allows the estimate Q to have a doubly robust property, making it an unbiased estimator if either the outcome or propensity score estimate is accurate. In **Appendix C.4**, we also consider DragonNet without the targeted regularization term.

5.3 Experiments and Results

Theory for techniques that incorporate estimates of both the propensity score and the potential outcomes often make them a preferred estimator over other categories of techniques. Hence, empirically, we aim to address the following questions to probe the applicability of different CATE estimation techniques in the context of neural networks:

- Do approaches that rely on both an estimate of the potential outcomes and the propensity score during training outperform approaches that rely only on the propensity score or modeling the outcome?
- How do our conclusions change when we introduce errors in the estimates of the propensity score?

We first describe our experimental set-up, including datasets and evaluation metrics that are used to measure performance. We then present results aimed to provide a better understanding of the relative performance of different methods.

5.3.1 Experimental Set-Up

To explore the gap between theory and practice for CATE estimation techniques, we investigate models across both synthetic and benchmark semi-synthetic datasets. Validating on real datasets is difficult due to the fundamental problem of causal inference, or the inability to observe both potential outcomes. Moreover, evaluations of CATE estimators on real datasets often rely on inaccurate proxy variables that must be estimated from the data. Errors in the problem set-up and evaluation may easily result in inaccurate takeaways, with the potential to lead to harm when applied to real data. Hence, as an important step towards the goal of real-world applications of these methods, we focus on existing synthetic and semi-synthetic benchmark datasets in which the counterfactual is available [182, 48, 141]; such datasets are designed to test practical aspects of CATE estimation. We consider testing these models under 1) the assumption of ground-truth propensity scores, and 2) the practical setting with estimated or noisy propensity scores to answer our primary research questions. For all

methods that require the use of estimated outcomes or propensity scores during training, though theory often requires these models to be built using different data than the downstream estimator, past work has found that using all training examples performs better in low-sample data [43, 145]. We hence follow this past work and use all training examples to build these estimators. In **Appendix C.3**, we confirm that this choice results in better empirical performance compared to a standard cross-fitting approach. Further training details, including model architecture, compute infrastructure and hyperparameter tuning can be found in **Appendix C.1**.

5.3.1.1 Datasets

Synthetic Dataset. We simulate a synthetic dataset in which we can control the degree of confounding and have access to ground-truth propensity scores. To induce confounding, we generate our dataset by imposing a direct relationship between the outcome and the propensity score, as in past work [71, 33]. Covariates that affect both the outcome and the propensity score induce confounding, leading to biased treatment effect estimates if not addressed. We begin by simulating a dataset with significant confounding bias. We simulate $\mathbf{x}_i \in \mathbb{R}^{10}$ for each individual, generated as correlated uniforms from a Gaussian copula, where the covariance matrix R is such that $R_{ij} = .5^{|i-j|} + .1I[i \neq j]$. The other quantities for individual i are simulated as follows:

$$
\begin{aligned}
&\mu(\mathbf{x}_i) = 5(2 + 0.5\sin(\pi \mathbf{x}_{i,1}) - 0.25\mathbf{x}_{i,2}^2 + 0.75\mathbf{x}_{i,3}\mathbf{x}_{i,9}),\\
&\tau(\mathbf{x}_i) = 1 + 2|\mathbf{x}_{i,4}| + \mathbf{x}_{i,10},\\
&e(\mathbf{x}_i) = 0.9\Lambda(1.2 - \gamma\mu(\mathbf{x}_i) + \eta_i),\\
&\eta_i \sim \mathcal{U}(0,1),\\
&t_i \sim Bernoulli(e(\mathbf{x}_i)),\\
&y_i = \mu(\mathbf{x}_i) + \tau(\mathbf{x}_i)t_i + \epsilon_i,\\
&\epsilon_i \sim \mathcal{N}(0,\sigma^2), \sigma^2 = \frac{\sigma_\tau}{2}
\end{aligned}
$$

where Λ is the logistic cumulative distribution function and γ controls the level of confounding in the dataset. The relationship between the baseline outcome μ and the propensity score e induces confounding. To measure confounding, we calculate the alignment, or the correlation between the observed outcome and the propensity score (i.e., $\rho(y, e)$) [48]. A strong absolute correlation indicates high confounding in the dataset.

We begin by simulating a dataset with significant confounding bias. We measure the level of confounding using *alignment*, or the correlation between the observed outcome and the propensity score (i.e., $\rho(y, e)$) [48]. A strong absolute correlation (i.e., high absolute

alignment) indicates high confounding in the dataset. In the synthetic dataset, we first set the magnitude of alignment to approximately 0.85 (i.e., $\gamma = 1$) in our experiments, resulting in a dataset with strong confounding. In ablation studies, we also analyze model performance as the level of confounding changes. We simulate $N = 1000$ examples for training and testing respectively and repeat the simulation 30 times with different random seeds and report results averaged over the replications (i.e., 30 datasets)

Semi-Synthetic Datasets. We also use data from the Collaborative Perinatal Project, a large longitudinal cohort study of pregnant women designed to study factors leading to developmental disorders [143]. As in past work, we use these data to simulate a twins study with the goal of estimating the impact of birth weight on a child's IQ [48]. First, we chose a set of covariates that serve as potential confounders. Next, we defined treatment assignment mechanisms and outcome functions using these data to test a variety of settings, resulting in 77 unique DGPs known as the **ACIC 2016** dataset. ACIC allows us to examine the performance of different techniques on a wide variety of relevant datasets with differing characteristics. Specifically, the datasets vary in their 1) degree of nonlinearity, 2) the percentage of treated individuals, 3) overlap, 4) alignment (i.e., confounding), 5) treatment effect heterogeneity, and 6) the magnitude of the treatment effect. For each DGP, we consider 30 simulations with different random seeds for generating treatments and outcomes. Each simulation within each DGP consists of 58 covariates. We train using 500 random examples to mimic a challenging small sample regime and further test the finite-sample nature of these estimators [3]. The ACIC dataset is available at https://github.com/vdorie/aciccomp.

Table 5.2: Synthetic dataset results when the **ground-truth propensity score** is available. This table shows the accuracy as measured by PEHE and the number of replications (out of 30) in which each model outperforms TARNet. The X-Learner outperforms all other techniques. Results in bold are a statistically significant improvement over TARNet.

Model	PEHE (SD) ↓	Improvements in PEHE ↑
TARNet	1.113 (0.201)	—
X-Learner	**0.711 (0.220)**	29
Weighting	**1.015 (0.161)**	20
Matching	**0.894 (0.163)**	24
R-Learner	**0.881 (0.139)**	26
DR-Learner	**0.714 (0.199)**	29
DragonNet	**1.011 (0.255)**	20

Table 5.3: Top performing models and average rankings on ACIC 2016 dataset across replications when using **ground-truth propensity scores**. Overall, the X-Learner obtains the best average performance. Many propensity score adjustment techniques, including DragonNet, perform poorly.

Model	# Top-Performing PEHE (/77) ↑	Average Ranking PEHE ↓
TARNet	2/77	4.66
X-Learner	29/77	**2.03**
Weighting	1/77	3.65
Matching	**35/77**	2.97
R-Learner	0/77	6.25
DR-Learner	10/77	3.14
DragonNet	0/77	5.30

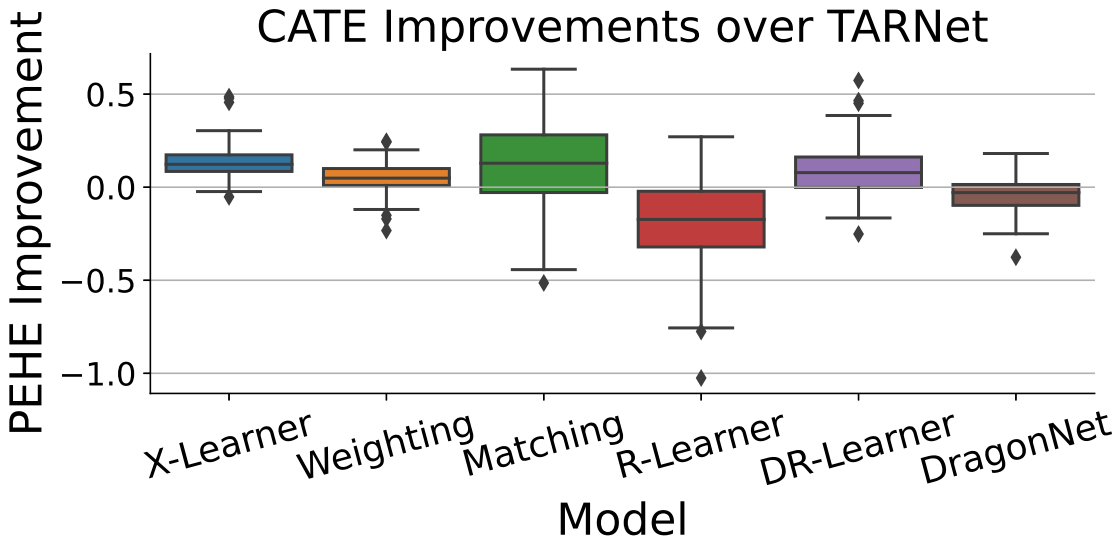

Figure 5.1: Model performance with **ground-truth propensity scores** relative to TARNet across all DGPs in the ACIC 2016 dataset. The X-Learner is the only model able to outperform TARNet in almost every DGP, though the DR-Learner performs well. DragonNet and the R-Learner fail to improve over TARNet in a majority of settings.

5.3.1.2 Evaluation Metrics

As outcomes are simulated for all datasets, as in past work [182, 74, 75, 92], we calculate the ground-truth performance of each model in terms of the **PEHE** defined in **Section 2.2.4**. In line with a majority of past work in the field, we are interested in evaluating models for their ability to accurately estimate pointwise CATEs [182, 215, 38]. However, in many

situations, accurate CATE estimates may not be the end goal; future works may consider how these methods perform with respect to other tasks, such as confidence interval creation and downstream use [48].

For the synthetic dataset, we average results over 30 replications and report the mean and standard deviation, and the number of replications in which a technique outperforms TARNet. We consider comparing all models to TARNet as it does not incorporate any extra adjustment beyond the confounders and simply models the outcome. By comparing to TARNet, we can better measure how extra adjustment techniques improve or fail to improve the model beyond this baseline. For the ACIC dataset, we also consider performance over the 30 replications for all DGPs. We are interested in understanding model performance across a wide variety of settings, and hence, do not consider averaging PEHE across all DGPs as in past work [185, 219, 88]. Instead, we report the number of DGPs in which each model was the top-performing algorithm and the average rank in performance for each model across all DGPs. We also visualize the relative performance of each method compared to TARNet across every DGP to understand their gain over the simplest baseline.

5.3.2 Results

Table 5.4: Synthetic dataset results when utilizing an **estimated propensity score** during training. All methods degrade, though the X-Learner remains the most robust and outperforms all other methods. Results in bold are statistically significant compared to TARNet.

Model	PEHE (SD) ↓	Improvements in PEHE ↑
TARNet	1.113 (0.201)	—
X-Learner	**0.713 (0.219)**	29
Weighting	1.073 (0.281)	21
Matching	1.615 (0.204)	0
R-Learner	1.285 (0.191)	7
DR-Learner	**0.774 (0.222)**	28
DragonNet	**1.011 (0.275)**	20

Results with Ground-Truth Propensity Scores. We first compare all techniques assuming access to ground-truth propensity scores. In the synthetic data, the X-learner, a technique that only uses models of the outcomes during training, outperforms almost all other techniques (PEHE: 0.711, SD: 0.220) (**Table 5.2**). This includes methods that incorporate estimates of the propensity score during training as well as techniques that use

both propensity score estimates and potential outcome estimates during training. The DR-Learner achieves similar performance to the X-Learner, outperforming all other techniques (PEHE: 0.714, SD: 0.199).

On the ACIC dataset, the X-Learner continues to outperform all techniques with an average ranking of 2.03 across all DGPs (**Table 5.3**). Among methods that incorporate estimates of both the propensity score and the potential outcomes, the DR-Learner continues to show strong performance compared to baselines (average ranking: 3.14). However, DragonNet and the R-Learner, despite their strong theoretical guarantees, perform very poorly with an average ranking lower than that of all other methods. In **Appendix C.4**, we find that DragonNet without targeted regularization performs better. However, the X-Learner still performs better.

Table 5.5: Top performing models and average rankings on ACIC 2016 using **estimated propensity scores** across 77 datasets. The performance of all propensity score adjustment techniques degrade, with X-Learner still remaining robust.

Model	# Top-Performing PEHE ↑	Average Ranking PEHE ↓
TARNet	3/77	4.03
X-Learner	**47/77**	**1.57**
Weighting	7/77	3.82
Matching	8/77	4.79
R-Learner	1/77	5.55
DR-Learner	10/77	3.45
DragonNet	1/77	4.79

To study results on the ACIC dataset further, we visualize the variability of each model's ability to outperform TARNet across DGPs (**Figure 5.1**). The X-Learner improves over all techniques, achieving the highest median performance with low variance across settings (Median improvement: 0.123, IQR: 0.085, 0.174). Compared to the next best-performing techniques (Matching median improvement: 0.129, IQR: -0.029, 0.281, DR-Learner median improvement: 0.078, IQR: -0.001, 0.162), the X-Learner provides the best trade-off between strong average performance and consistent performance across settings. Other techniques have high variability in terms of their ability to outperform TARNet, with most techniques unable to consistently do so and some like the R-Learner and DragonNet consistently performing worse. **Figure 5.2** shows that the performance of many methods, as measured by improvement over TARNet, is positively correlated with the level of confounding present in the dataset. In the synthetic dataset, TARNet outperforms or matches all models except the X-Learner and the DR-Learner at confounding levels below an alignment of 0.8. Moreover,

matching performs best at high levels of confounding but can be outperformed substantially by TARNet at lower levels, explaining the high variability of the performance of matching on the ACIC dataset. In the ACIC dataset, the ability of most methods to improve over TARNet has a positive correlation with the level of confounding. However, the X-Learner consistently outperforms all other methods across levels of confounding. Moreover, at lower levels of confounding, simpler methods like TARNet may be preferable over many more complex propensity score adjustment techniques, even with ground-truth propensity scores.

Results with Incorrect Propensity Scores. In the previous section, we assumed access to ground-truth propensity scores. However, in practice, ground-truth propensity scores are rarely available. In our final set of experiments, we examine the performance of these approaches when propensity scores are noisy or estimated. Note that TARNet and DragonNet are unaffected by these changes, as they do not use explicit estimates of the propensity score.

First, we consider when propensity score estimates are noisy yet consistent and unbiased (i.e., in expectation, the propensity scores are correct). To do so, we perturb ground-truth propensity scores by adding zero-mean Gaussian noise with increasing standard deviation. These noised propensity scores are clipped between to be within 0 and 1 and are then incorporated into the relevant model training and evaluation schemes for different approaches.

On the synthetic dataset, when the standard deviation of noise is small, most methods that rely on propensity scores during training still perform well (**Figure 5.3**). However, as the noise added to the propensity scores increases, most techniques degrade in performance, eventually performing worse than TARNet. The DR-Learner degrades the most despite the use of potential outcome estimates. This can be explained by the fact that if the nuisance models are both misspecified, The DR-Learner no longer necessarily holds strong theoretical guarantees. The X-Learner remains robust, consistently outperforming all techniques and only performing slightly worse as the level of propensity score error increases relative to other techniques. We note that the performance of the X-Learner does decrease with worse propensity scores in this setting. When the propensity scores are flipped, the X-Learner performance degrades from a PEHE of 0.711 to 0.761. However, even in this situation, the X-Learner is more robust than other techniques. In the ACIC dataset, the results are similar. The PEHE of all methods using propensity scores during training increases as the level of noise added to the propensity score increases, with the DR-Learner degrading the most. The X-Learner remains robust as the propensity score is only used to weight the learned CATE estimators at inference time.

Second, we consider the performance of different techniques when propensity scores must be estimated and may be biased. We estimate propensity scores using logistic regression,

a widely used technique in past work [15, 128]. Though there are many techniques for estimating the propensity score, we focus on a simple and widely used technique, to examine the implications of this choice. The results remain similar to the previous setting in both the synthetic dataset and ACIC datasets, where the performance of all techniques besides X-Learner degrade. In the synthetic dataset, the only method that uses explicit estimates of the propensity score during training that is able to consistently outperform TARNet is the DR-Learner, but even its performance degrades compared to when it is given access to ground-truth propensity scores (PEHE: 0.774, SD: 0.222 vs. PEHE: 0.714, SD: 0.199) (**Table 5.4**). Meanwhile, the X-Learner remains robust, outperforming all techniques (PEHE: 0.713, SD: 0.219). On the ACIC dataset, the same results hold. The X-Learner outperforms all other techniques (average ranking: 1.57), with the performance of all propensity score adjustment techniques degrading substantially (DR-Learner: average ranking 3.45 vs. 3.14) (**Table 5.5** and **Figure 5.4**).

In the presence of inaccurate propensity scores, the performance of many techniques degrades substantially. Recent work has shown the effect of propensity score errors may have a near-negligible effect on downstream CATE estimation when an estimate of the outcome is also incorporated [40, 56, 142, 105]. However, our empirical results run in direct contrast to this and show how errors in the propensity score may cause extreme degradation in CATE estimation performance. Hence, even in datasets with a high degree of confounding, without access to ground-truth propensity scores, many more complicated techniques may not outperform models that simply rely on estimates of the outcome, such as the X-Learner.

5.4 Conclusion and Discussion

There exist many popular techniques for adjusting for confounding when estimating CATEs. Techniques that incorporate estimates of the propensity score and potential outcomes during training are popular and theoretically strong techniques that are often favored over other approaches. However, theoretical results for these techniques often do not consider practical finite-sample performance, limiting their ability to provide guidance in model selection. In this chapter, we provided an extensive comparison of a wide variety of CATE estimation techniques across a multitude of different settings using neural networks as base learners. Our empirical analysis led to important findings that should be considered when building and using CATE estimation techniques.

First, the X-Learner was able to consistently outperform all techniques across a multitude of different settings, even when these approaches were given access to ground-truth propensity scores during training. This includes strong baselines like the DR-Learner that in-

corporate both estimates of the propensity score and the potential outcomes during training. This points to the sufficiency of outcome-based modeling approaches in practical settings. Second, we demonstrated the sensitivity of different CATE estimation techniques to errors in propensity score estimates. We found that such errors can result in a non-negligible decrease in CATE estimation performance, even for doubly robust techniques like the DR-Learner. In these settings, the X-Learner, or even simpler algorithms like TARNet, may be preferable over these more complicated techniques. Our work suggests that outcome-based modeling techniques may be a better choice in many practical settings for CATE estimation, especially when confounding levels are low and propensity scores cannot be estimated accurately.

As with all empirical analyses, our work is not without limitations. We examine a limited set of neural network architectures and CATE estimation techniques. However, we consider a well-studied popular architecture from past work and popular CATE estimation approaches in the literature. Our results are an important case study that future research may build on to test different settings and approaches.

Our work addresses an important gap between theory and practice in CATE estimation, demonstrating the importance of rigorously evaluating techniques across a variety of settings to complement theoretical results. Our findings represent important future steps and practical considerations when learning CATEs using neural networks. Identifying and empirically investigating assumptions derived from theoretical results is critical for practical progress across many fields in machine learning [122, 199]. Given these findings, it remains imperative to further investigate popular approaches for CATE estimation, towards the goal of real use and impact.

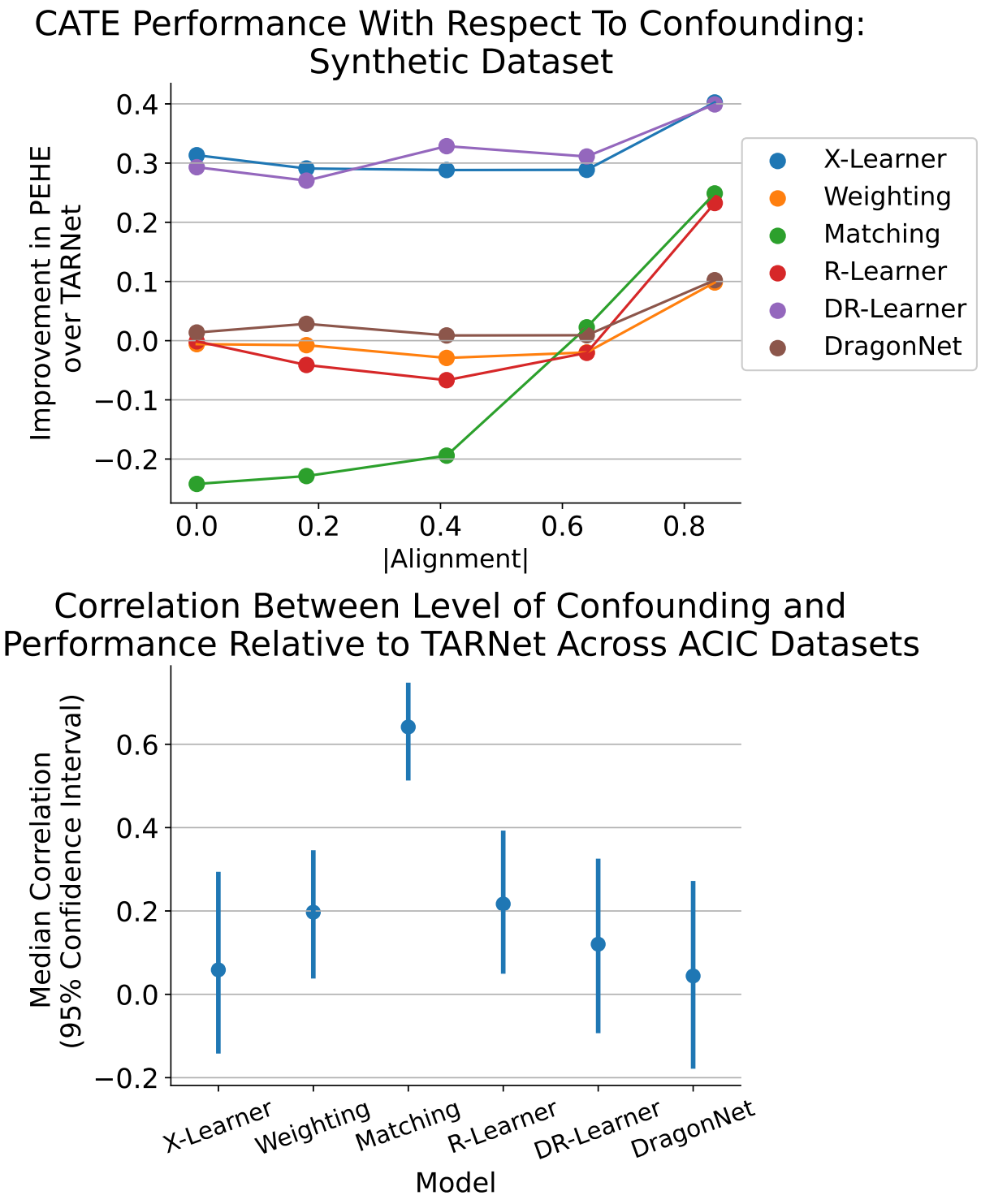

Figure 5.2: The ability to improve over TARNet varies as confounding is changed in the synthetic dataset (top), with many methods unable to outperform TARNet at lower levels of confounding. There exists a significant correlation between the level of confounding and the performance compared to TARNet for most methods in the ACIC dataset (bottom).

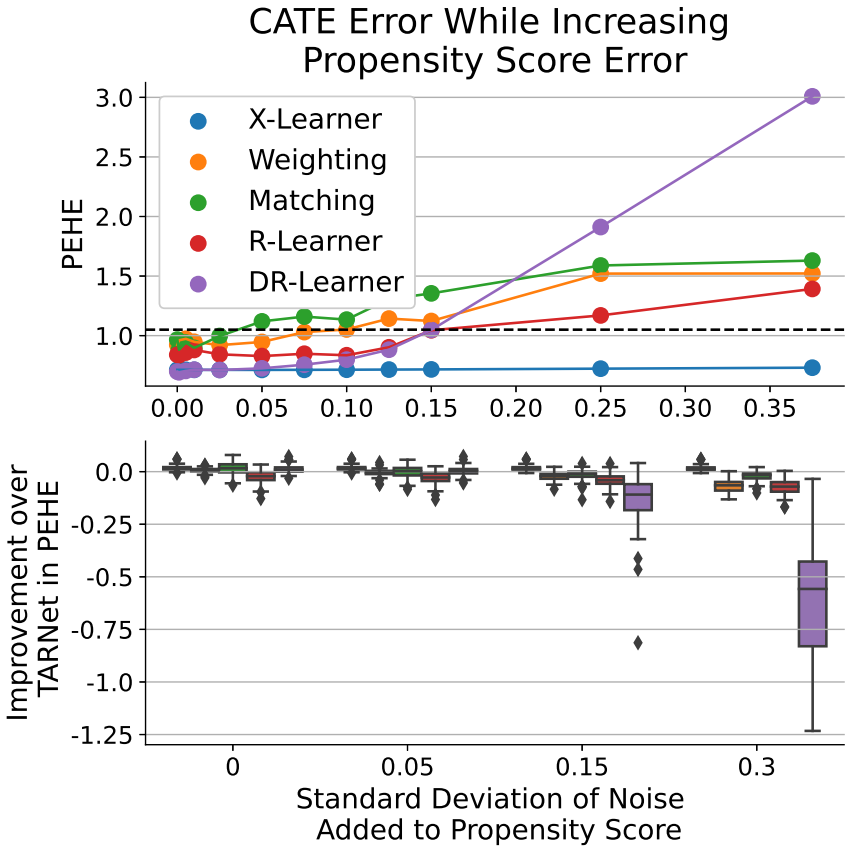

Figure 5.3: CATE error across models as the propensity score is artificially noised on (top) the synthetic dataset and (bottom) the ACIC 2016 dataset, over all DGPs. The black-dashed line shows TARNet performance on the synthetic dataset. Once propensity scores are sufficiently noisy, all methods are outperformed by covariate adjustment approaches, including both TARNet and the X-Learner.

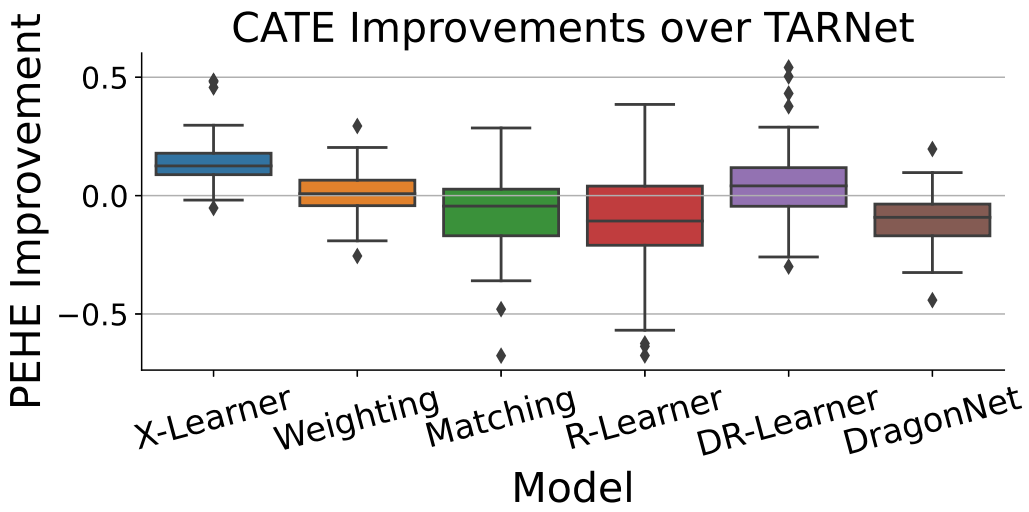

Figure 5.4: CATE improvements of CATE estimation models with **estimated propensity scores** over TARNet across all DGPs. All methods deteriorate, though the X-Learner dominates over all methods.

CHAPTER 6

Mismatch in Sepsis Risk Stratification and Clinical Needs

6.1 Introduction

Finally, we consider how mismatches between machine learning research and real clinical needs manifest through a case study of existing clinical tools for the problem of sepsis risk stratification. Sepsis contributes to approximately one out of three in-hospital deaths in the US [1, 158, 204, 54, 175, 121]. Timely identification and treatment of patients with sepsis can lead to significant improvements in mortality rates among hospitalized patients [188, 53, 117, 160, 60, 47, 162]. To enhance clinical decision-making, recent work has focused on developing predictive models that use electronic health record (EHR) data to identify patients at risk of sepsis [77, 208, 44, 207, 106]. For example, the Epic sepsis model (ESM) is one of the most widely implemented systems in US hospitals [164, 127]. Similar to the ESM, the majority of risk stratification tools aim to identify patients at high risk of developing sepsis prior to the sepsis criteria being met [2, 178, 184, 183]. In this chapter, we first study the mismatch between how these models are currently evaluated and their ultimate goal of augmenting clinical intuition. From here, we take a step back to understand whether the use of current risk stratification tools aligns with the goal of improving patient outcomes.

6.2 Mismatch Between Evaluation of Sepsis Risk Stratification Tools and Clinical Utility

First, we focus on potential limitations of the evaluation of current sepsis risk stratification tools. Models such as the ESM make predictions throughout an individual's hospitalization, incorporating relevant changes in a patient's health status based on the contents of the EHR. Discriminative metrics, such as the area under the receiver operating characteristic

curve (AUROC), are commonly used to assess model performance [208, 125, 174]. Typically, AUROCs are calculated at a hospitalization-level using predictions before a patient meets sepsis definitions with the goal of evaluating the model's ability to predict sepsis before it occurs [208, 144]. However, clinicians may recognize or begin to treat sepsis well before it is definitively diagnosed. Consequently, predictions occurring after treatment may not be as clinically useful as those before recognition. This phenomenon, referred to as "label leakage," may be exacerbated in models that include treatments (e.g., antibiotics) as predictors because the accuracy of such models may largely be derived from predictions made after clinical recognition [205, 61]. Though this may lead to greater apparent performance, these predictions do not provide clinicians with new information and instead may contribute to alert fatigue [34]. In such cases, many patients correctly identified by the model as high-risk have already been identified by healthcare practitioners [168]. In contrast, a model that can identify high-risk patients with sepsis before a clinician recognizes signs or symptoms of sepsis could enable the more timely delivery of care.

We study the mismatch between how current sepsis risk stratification models are evaluated and their ultimate goal of adding to clinical intuition. While it is difficult to retrospectively identify which patients a clinician would have otherwise missed, we can evaluate the accuracy of predictions in advance of indicators of sepsis treatment, where such indicators serve as proxies for clinical recognition. We introduce and apply a new sepsis model evaluation framework that incorporates the timing of various indicators of sepsis treatment and use it to evaluate the ESM to understand the performance of the model with respect to clinical recognition of sepsis. Our analysis highlights the gap between existing evaluation schemes and accurately measures the utility of models for augmenting downstream clinical decision-making.

6.2.1 Methods

Study Cohort. Our retrospective cohort included adult inpatients admitted to the University of Michigan's academic medical center, Michigan Medicine (MM) between October 2018 and December 2020. We included all hospitalizations in this time period for evaluation except hospitalizations from psychiatric and rehabilitation units. This study was approved by the institutional review board (IRB) at Michigan Medicine (HUM: 00176141), and the need for consent was waived as there was minimal risk to participants.

Definition of Sepsis, Onset Time, and Sepsis Treatments. Sepsis was defined based on a composite definition of multiple criteria. The composite definition was based on meeting one of the following two definitions: 1) the clinical surveillance definition defined

by the Center for Disease Control and Prevention (CDC) [159, 157], or 2) the Centers for Medicare and Medicaid Services (CMS) definition, corresponding to meeting 2 criteria for systemic inflammatory response syndrome (SIRS) and 1 criterion for organ dysfunction within 6 hours of one another (i.e., SEP-1) [198, 95]. For hospitalizations meeting the CMS definition, the time of meeting the sepsis criteria was defined as the later time of meeting the SIRS criteria or the organ dysfunction criteria. For hospitalizations that did not meet the CMS definition, the time of meeting the sepsis criteria was defined as the first time in which the CDC definition was met. Through electronic data capture, the timing and compliance of the ordering and administration of indicators for sepsis treatment were measured with respect to the time at which sepsis criteria were met [117, 60]. Initial validation of the data capture was completed manually by a team of clinical analysts and engineers to ensure nearly perfect accuracy in identifying relevant treatment indicators. We included treatment and diagnostic orders as indicating the initiation of a treatment plan: fluids, antibiotics, lactate measurement (captured from both order sets and individual orders), and blood culture.

The Epic Sepsis Model. The ESM is a sepsis risk model developed by Epic Systems Corporation, Verona, Wisconsin [208, 193]. The ESM uses data recorded within the EHR to make predictions every 20 minutes during a hospitalization. The ESM is a penalized logistic regression model that outputs a continuous score between 0 and 100, where 0 represents the lowest possible risk and 100 represents the highest. We considered individuals with ESM scores before the first of 1) meeting the sepsis criteria, 2) ordering of any indicator for treatment of sepsis, or 3) death or discharge.

Evaluation Framework. We evaluated model predictions at key time points during the hospitalization to understand the performance of the model in relation to clinical recognition of sepsis (**Figure 6.1**). To mimic a situation in which a clinician has not yet recognized sepsis and initiated treatment, we used predictions preceding when sepsis criteria were met and the first indicator of treatment. We compared the predictive performance resulting from the evaluation above with that achieved when evaluating using only predictions before the time that the sepsis criteria were met. In addition, as an upper bound on performance, we evaluated model performance using predictions up until discharge, including predictions made after sepsis criteria may have been met. We calculate a hospitalization-level AUROC, where an AUROC of 0.5 means a model's performance is no better than random. To calculate a hospitalization-level AUROC, we take the maximum ESM score as the hospitalization-level score for predictions before each evaluation time point separately for each hospitalization. This evaluation mimics how the ESM would generate alerts in a real clinical setting, where an alert would be fired for a particular hospitalization if the threshold was ever exceeded up to the time point of interest [208, 144]. We estimated the 95% confidence interval of

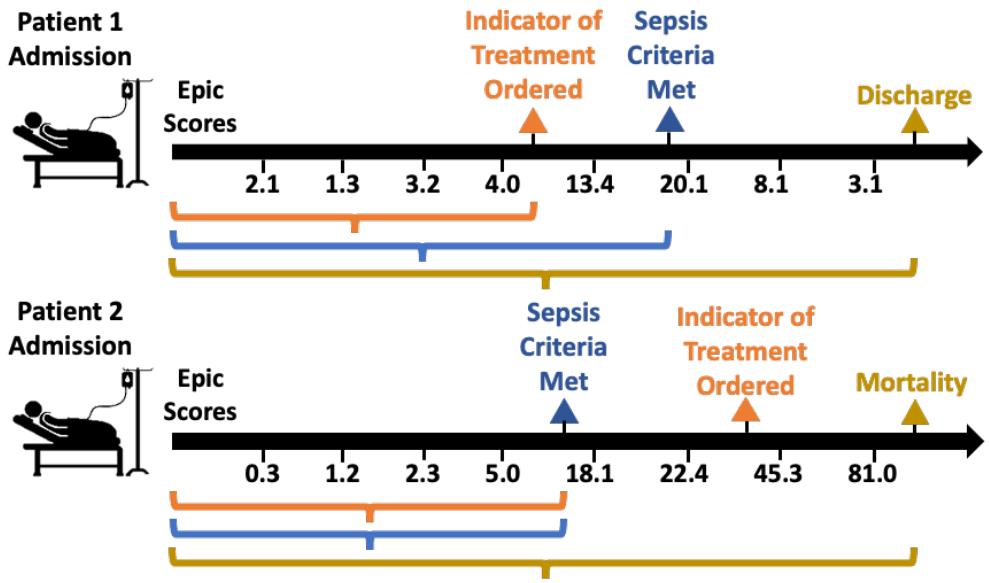

Figure 6.1: Overview of different evaluation schemes. In Patient 1, indicators of treatment for sepsis occur before sepsis criteria is met. In Patient 2, sepsis criteria is met before any treatment indication. If the model is relying on treatment indicators, then in the case of Patient 1, ESM model accuracy should decrease if data collected after the initiation of treatment are excluded. However, for Patient 2, ESM model accuracy should not change because no treatments were ordered before the time the sepsis criteria was met. For both patients, the highest ESM model accuracy should occur when using all data up to the time of discharge.

the AUROC with 1,000 bootstrap samples. We also measured the positive predictive value (PPV) and sensitivity of the ESM at a score threshold of 6, which is currently used to generate alerts at MM.

In a sensitivity analysis, we separately evaluated with respect to each treatment indicator (antibiotics, lactate measurement, fluids, blood culture) to better understand which treatment indicators drove changes in performance. In addition, we evaluated with respect to diagnostic orders (lactate and blood culture collection) and treatment orders (antibiotic and fluid administration) separately.

Secondary Analyses: Adjusting for the Amount of Data Available. Evaluating sepsis prediction models based on the timing of indicators for treatment could exclude significant portions of a patient's hospitalization, thereby reducing the amount of available data for making predictions. In many cases, clinicians may order treatments well in advance of sepsis criteria being met. To measure the amount of data available at different points of evaluation, we quantify the number of laboratory and medication orders for admissions in

which the individual met the sepsis criteria at each time-point of evaluation. To control for the amount of data available to the algorithm, we stratify the number of orders available at treatment initiation into quintiles, and we evaluate the ESM within each stratum. We compare ESM performance when evaluating with respect to the time of the first indicator for treatment and the time of the sepsis criteria being met. Across all quintiles, we keep all individuals without sepsis as negative examples to calculate the AUROC.

6.2.2 Experiments and Results

Through our experiments, we aim to answer the following questions:

1. Do clinicians recognize and treat sepsis prior to sepsis criteria being met?
2. How does the performance of models differ when evaluating with respect to varying levels of clinical recognition?
3. Are differences in performance among different evaluation schemes solely due to the amount of data available to the model?

Population Characteristics. We identified 77,582 hospitalizations that met our inclusion/exclusion criteria for the study cohort. Of these hospitalizations, sepsis occurred in 3,766 (4.9%). A total of 3,538 (93.9%) hospitalizations with sepsis had some indicator of sepsis treatment. Over 70% of the hospitalizations with sepsis received orders for antibiotics (76.4%), blood culture (72.4%), or lactate measurement (77.6%) as part of a treatment plan for sepsis, while only 29% were ordered some level of fluids for sepsis. Over 45% of sepsis hospitalizations had antibiotics, blood culture, or lactate measurements ordered before the time of sepsis, with median lead times of 55 minutes, 46 minutes, and 43 minutes before sepsis criteria were met respectively (**Figure 6.2**). Treatment indicators preceded the time of meeting the sepsis criteria in 3,193 (84.8%) of hospitalizations. Lactate was the first treatment indicator ordered in 47.1% of hospitalizations, followed by antibiotics (23.7%) and blood cultures (20.0%).

Primary Analysis Evaluation With Respect to Varying Degrees of Clinical Recognition. Using all predictions up until discharge during a hospitalization, the model achieved an AUROC of 0.87 (95% CI: 0.86-0.87), a PPV of 16% (95% CI: 16%-17%), and a sensitivity of 79% (95% CI: 78%-80%). Using only predictions before meeting the sepsis criteria, the ESM model had an AUROC of 0.62 (95% CI: 0.61- 0.63), a PPV of 8% (95% CI: 8%-9%), and a sensitivity of 38% (95% CI: 36%-39%). Further restricting to predictions made before treatment indicators, performance decreased, with an AUROC of 0.47 (95%

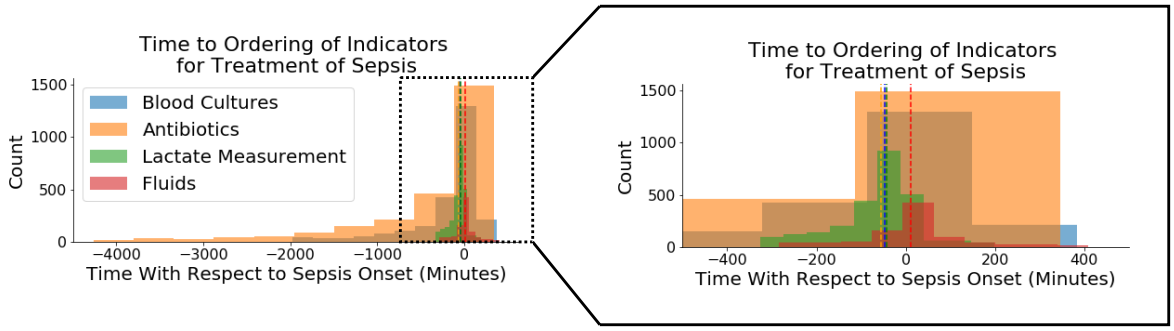

Figure 6.2: Temporal distribution of indicators of treatment with respect to sepsis criteria time. The dashed vertical bars represent the median time for each treatment. Antibiotics, blood culture collections, and lactate measurements are ordered substantially before the time of sepsis. Nearly half of the population has orders for lactate measurement, antibiotics, or blood cultures before the onset of sepsis.

CI: 0.46-0.48), a PPV of 5% (95% CI: 4%-5%), and a sensitivity of 20% (95% CI: 19%-22%). Performance dropped most when predictions were restricted to before the time of blood culture orders (AUROC: 0.53, 95% CI: 0.52-0.54, PPV: 5.9%), and dropped the least when predictions were restricted to before the time of fluid orders (AUROC: 0.61, 95% CI: 0.60-0.620, PPV: 8.0%) (**Figure 6.3**). When evaluating at the first diagnostic order (i.e., lactate and blood culture collection), the ESM achieved an AUROC of 0.9 (95% CI: 0.48, 0.50). Meanwhile, when evaluating at the first treatment order (i.e., antibiotic and fluid administration), the ESM achieved an AUROC of 0.55 (95% CI: 0.54, 0.56).

Secondary Analyses: Adjusting for the Amount of Data. For a majority of cases, treatment indicators preceded when sepsis criteria was met (84.8%). At treatment indicator time, individuals had on average significantly fewer orders compared to when sepsis criteria is met (median count 22 [IQR: 9-92] vs. 79 [IQR: 28-187]). Adjusting for the amount of data available to the algorithm, the ESM model consistently performed worse when evaluating before treatment indicator time rather than before sepsis criteria meeting time across all levels of available data (**Figure 6.4**). However, the gap decreased as more data become available to the ESM model, as measured by the number of orders.

Overall, clinicians tended to order treatments for sepsis before sepsis criteria were met. Moreover, the performance of the ESM dropped significantly when evaluating prior to clinical recognition of sepsis compared to the standard evaluation using data prior to sepsis criteria being met. This trend remained true even when adjusting for the amount of data given to the model. Our analysis points to an important mismatch between how the utility of sepsis risk stratification tools is currently measured and how these tools are used downstream to augment clinical decision-making.

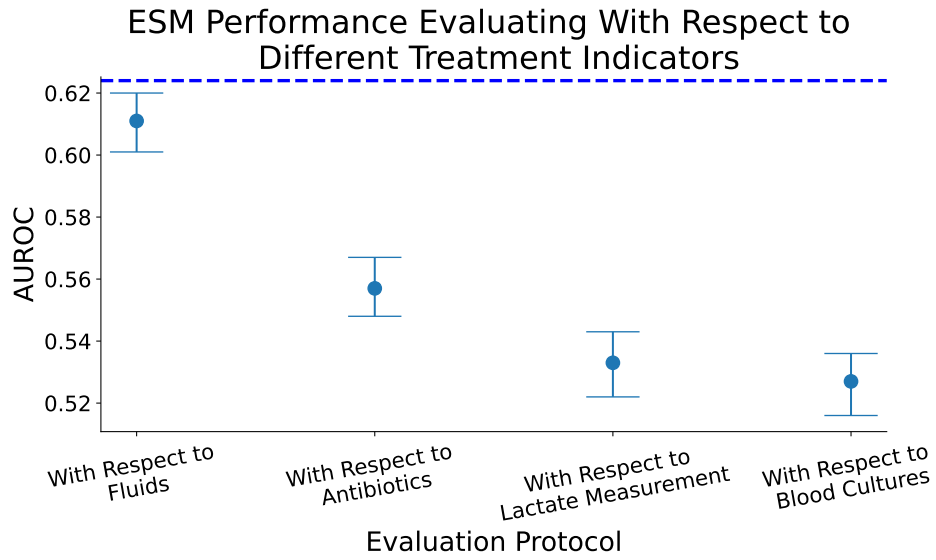

Figure 6.3: Evaluating the ESM with respect to different treatments. Evaluating the ESM with respect to different treatments. We visualize the performance with 95% confidence intervals for each evaluation. The blue dashed line denotes the ESM performance with respect to sepsis criteria time. The model performance drops the most when evaluating using predictions before the time of blood culture orders, achieving nearly random performance. Meanwhile, model performance only drops slightly when using predictions before orders for fluids.

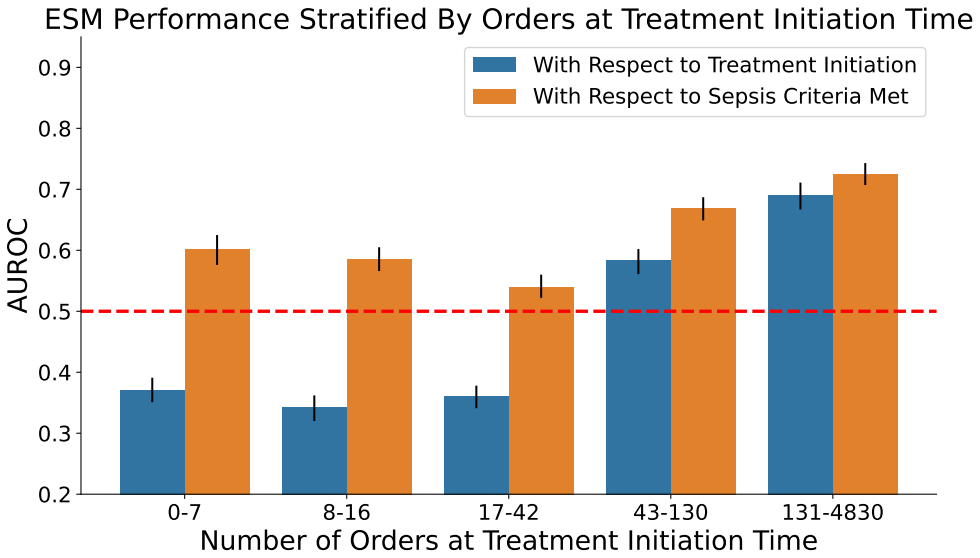

Figure 6.4: Evaluating the ESM with respect to different treatments. We visualize the performance with 95% confidence intervals for each evaluation. The blue dashed line denotes the ESM performance with respect to sepsis criteria time. The model performance drops the most when evaluating using predictions before the time of blood culture orders, achieving close to random performance. Meanwhile, the model performance only drops slightly when using predictions before fluid ordering.

6.3 Mismatch Between Estimating Risk of Sepsis and Improving Patient Outcomes

In the previous section, we studied the mismatch between how current sepsis risk stratification tools are evaluated and how they may best augment clinical intuition in practice. Our new evaluation procedure helped highlight the limitations of existing risk stratification tools when the goal is to predict sepsis. In this section, we take a step back to model development and study the mismatch between the objective of current risk stratification approaches and the goal of improving patient outcomes.

To date, the majority of work in patient risk stratification has focused on approaches that identify individuals at risk of developing a disease and often overlook the heterogeneous effects of the disease on patient outcomes [44, 77, 106, 136, 207, 208]. Interventions can then be allocated to those most likely to develop sepsis as identified by the model. When the goal is to improve patient outcomes, this approach assumes that those at risk of developing sepsis are also most likely to experience severe disease. However, the validity of this assumption has not been well-studied. Concretely, there may be patients who are likely to develop a disease but who are unlikely to suffer complications or die from it. Interventions targeting

these individuals, at the cost of delaying treatment to those who may be less likely to contract a disease but more likely to suffer complications or die, may be detrimental to the goal of improving patient outcomes. In this study, we probe the potential shortcomings of the objective of existing sepsis risk stratification tools and study the importance of considering disease severity in the context of risk stratification tools for sepsis. We focus on estimating the effect of sepsis on mortality within two large clinical cohorts and compare this with the estimated risk of developing sepsis at the level of the individual. Our analyses uncover significant heterogeneity in disease severity that only weakly correlates with the risk of developing sepsis. Beyond sepsis, this highlights the importance of accounting for downstream heterogeneity when building patient risk stratification models to guide interventions and further highlights the mismatch between existing risk stratification tools focused on predicting the likelihood of sepsis and the ultimate goal of improving patient outcomes.

6.3.1 Methods

6.3.1.1 Problem Set-Up and Cohort Definition

Study Cohorts. We use two retrospective cohorts. The first includes adult patients admitted to Michigan Medicine the hospital affiliated with the University of Michigan (U-M) between 2016 and 2020. In our primary analysis, we focus on only admissions to the ICU. The second cohort includes adult patients admitted to the ICU at Beth Israel Deaconess Medical Center between 2008 and 2012 (BIDMC) [94]. For both cohorts, we excluded admissions in which a suspected sepsis infection occurred prior to ICU admission, after ICU discharge, or within 1 hour of the model data collection, and admissions with missing data as defined by missing chart events or missing admission or discharge times [93, 136, 31]. In a secondary analysis involving the U-M dataset, we do not limit ourselves to ICU-only admissions and include in-patients across the entire hospital for evaluation (see **Appendix D.1.6**). The use of the U-M dataset for this study was approved by the institutional review board at the University of Michigan (HUM: 00176141). The BIDMC cohort is publicly available through Physionet [94].

Outcome Definitions. In the U-M cohort, we define sepsis similarly as before and use a composite of meeting either 1) the clinical surveillance definition created by the Center for Disease Control and Prevention (CDC), or 2) the Centers for Medicare and Medicaid Services (CMS) definition, with onset defined similarly as in the previous section [95, 157, 198]. Within the BIDMC cohort, we could not obtain accurate information relevant to the CDC surveillance definition. Thus, in line with past work, we used a pragmatic definition based on the Sepsis-3 criteria, defining onset time by identifying the acquisition of a body fluid

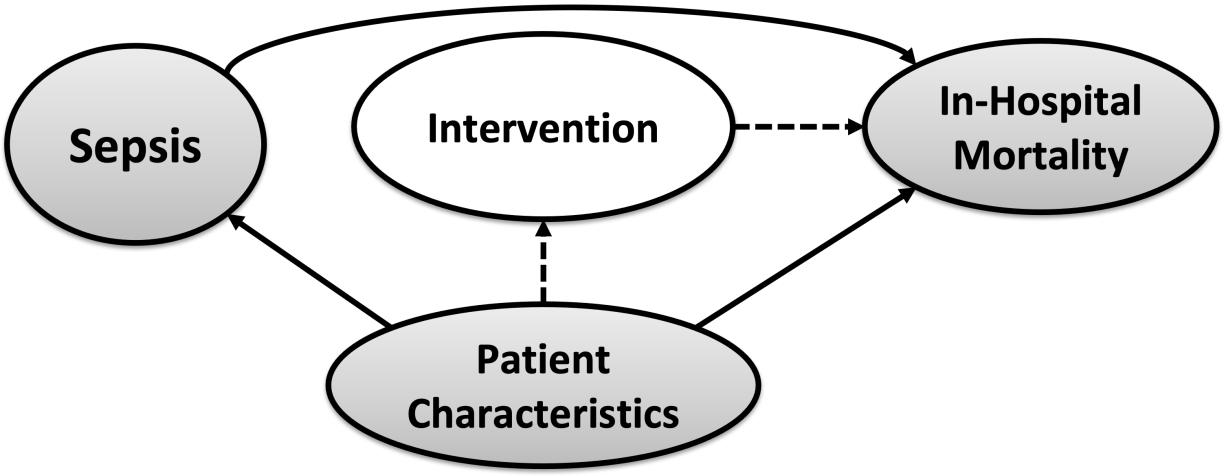

Figure 6.5: The assumed causal graph for our work. The dashed lines represent causal relationships for the treatment that is not currently captured in the data Patient characteristics affect the likelihood of sepsis and mortality, all of which are fully observed in our data. Sepsis also affects the likelihood of mortality. Finally, there exists a potentially novel intervention currently not observed in the data. Our goal is to understand how to allocate interventions to patients to reduce the overall mortality rate.

culture temporally contiguous to the administration of antibiotics [93, 136]. For both cohorts, in-hospital mortality was identified by utilizing discharge information.

Feature Extraction. For all patient admissions, we collect demographics, vital sign measurements, laboratory test results, and nursing score information, such as Glasgow coma scores and sedation information, throughout the hospitalization (see details in **Appendix D.1.2**). We include features that could be potential confounders between the development of sepsis and the likelihood of mortality to ensure the identifiability of causal effects. We do not include features that are treatments after sepsis has been recognized, such as the use of antibiotics or infrequently collected laboratory tests (i.e., laboratory tests collected in less than 40% of encounters in the U-M cohort). Despite best efforts to utilize similar features in the U-M and BIDMC cohorts, feature categories differed slightly. The full list of features considered can be found in **Appendix D.1.2**. All EHR data was preprocessed separately for each cohort using FIDDLE with the default settings [192].

6.3.1.2 Model Development and Evaluation: Estimating Sepsis Risk and Sepsis Severity

Overview. We study a scenario in which a patient's characteristics affect both their likelihood of developing sepsis as well as their likelihood of death during the current hospitalization. Developing sepsis has a direct effect on mortality, but this effect may be heterogeneous

among patients. While past work has focused on estimating treatment benefits [85, 49, 129], we assume a setting in which we aim to target some novel intervention not present in the available data (e.g., additional monitoring) (**Figure 6.5**). Given this causal model, we aim to estimate an individual's risk of developing sepsis and an individual's risk of severe sepsis as measured by the increase in the likelihood of in-hospital mortality. To estimate sepsis risk, we build a machine-learning model that maps an individual's characteristics throughout their admission to a probability estimate of the likelihood of developing sepsis. To estimate sepsis severity, we use causal inference techniques to map an individual's characteristics to an estimate of the effect of developing sepsis on their risk of mortality. This effect estimate is between -1 (i.e., developing sepsis decreased the likelihood of mortality from 100% to 0%) and 1 (i.e., developing sepsis increased the likelihood of mortality from 0% to 100%).

Model Development: Estimating the Risk of Developing Sepsis (Sepsis Risk). To estimate an individual's risk of sepsis, we train an ensemble of XGBoost models for each cohort to predict the likelihood of sepsis at every hour in an admission in line with past work [100, 144]. We use the XGBoost model as it has achieved strong performance in past work for predicting sepsis as it flexibly captures non-linear relationships between the patient's features and the development of sepsis [218, 213]. XGBoost models are preferred over other techniques due to their simplicity and ability to be combined with existing interpretability and explainability techniques, while maintaining strong performance [81, 221]. We split the data for each cohort into development and evaluation cohorts. In the U-M cohort, for model development, we use all inpatients from January 2016 to October 2018. In our evaluation cohort, we focus on only admissions to the ICU from October 2018 to December 2020. For BIDMC, we randomly split the data such that 70% of the admissions are used to develop models in this dataset, and the remaining 30% are used for evaluation. During training, we use a single window of one hour of data randomly sampled for each hospitalization, only including windows prior to the first of death, discharge, or sepsis onset. We repeat this process 50 times, leading to 50 different XGBoost models. We choose an ensemble in this manner to ensure we can learn from multiple windows during a patient's admission while reducing variance and increasing the accuracy of the model [152]. At inference time, the outputs of each model are averaged to create a final prediction for each window. We select hyperparameters for each model using 5-fold cross-validation, maximizing the AUROC for each of the 50 randomly sampled training cohorts (see **Appendix D.1.3** for detail).

Applied to the held-out evaluation cohort, we separately evaluate the sepsis risk model for each cohort in terms of the AUROC for predicting sepsis at the hospital admission level. Here, AUROC was calculated at the hospital admission level, taking the maximum of each score among the hourly windows [100, 208, 144]. We estimate the 95% confidence interval

with 500 bootstrap samples.

Model Development: Estimating Effect of Sepsis on Mortality (Sepsis Severity). We continue with making the assumptions from **Section 2.2.2** necessary for causal effect estimation. Given these assumptions, to estimate the effect of sepsis on mortality, based on best practices, we use a multitude of causal inference techniques. We apply these techniques independently to both cohorts, splitting the data into development and held-out evaluation cohorts as above. We train each model to map patient features from each one-hour data window to the effect of developing sepsis on the likelihood of mortality. We use three popular causal inference algorithms. First, we train an S-Learner, which predicts in-hospital mortality using both the covariates and the observed sepsis label as an extra covariate [112]. We use this model to estimate the effect of sepsis on mortality for a particular hospitalization by taking the difference between the model outputs when using the covariates with the sepsis label set to 1 and 0 respectively. Next, we also use the X-Learner and the DR-Learner due to their strong performance in **Chapter 5**. Here, we split the training data in half and used separate data for the first and second stages of these models to prevent overfitting. All models are an ensemble of XGBoost models, and hyperparameter selection was based on 5-fold cross-validation to maximize performance (see **Appendix D.1.3** for details).

Due to the lack of ground truth, we use an approximate metric to evaluate the causal estimates on the held-out evaluation cohorts. We perform a global null analysis, separately training causal inference techniques using random treatments in both the treatment and control groups [211, 210]. In such a situation, the ground-truth treatment effect should be 0 as the treatment is random. We evaluate each model by calculating the mean squared error between each estimated treatment effect and 0 across all estimates for all admissions and hourly windows in each evaluation set [211, 210]. We estimate the 95% confidence interval with 500 bootstrap samples. As this metric is simply approximate, it does not perfectly evaluate the accuracy of different causal inference models. For completeness, we run our remaining analyses using all causal effect estimation models and ensure that our main findings consistently hold regardless of the model used, as suggested by past work [210, 112]. Finally, we also evaluate the S-Learner's ability to accurately estimate mortality within both the septic and non-septic populations in terms of the AUROC on the evaluation set. Note that, by construction, this is impossible to measure for the X-Learner and the DR-Learner.

6.3.1.3 Statistical Analysis

Heterogeneity in Sepsis Severity. To examine the effect of sepsis on mortality, we first visualize all estimated treatment effects for each hospital admission and report the median

on each evaluation set. To measure heterogeneity, we calculate entropy by creating a discrete probability distribution by binning the learned estimated effects into 20 equal-sized intervals between -1 to 1. By binning the estimated effects in this manner, we can understand the variability of sepsis severity across 10% thresholds. In this setup, the maximum entropy possible is 3 (i.e., the entropy when there is an equal proportion of examples within each interval), while the minimum entropy is 0. As we make effect predictions at an hourly level, we aggregate all predictions by calculating the mean estimate for each admission across all windows.

Correlation Between Sepsis Risk and Sepsis Severity. To understand the relationship between the likelihood of developing sepsis and the effect of sepsis on mortality, we calculate the Spearman's correlation between the two estimated values. To estimate the relationship between these variables at a per-admission level, we again aggregate all predictions by calculating the mean estimate for each admission across all windows. We estimate the 95% confidence interval with 500 bootstrap samples of this aggregated dataset. To further visualize the relationship between sepsis risk and sepsis severity, we plot the mean and 95% confidence interval of the estimated effect of sepsis on mortality for each quintile of the estimated risk of sepsis separately across all windows for each cohort. We also visualize the empirical distributions of estimated effects for high-risk and low-risk windows respectively, where high-risk is defined as the top 20% of estimated risk of sepsis among all windows for each dataset, where 20% is chosen to match the alert rate of existing sepsis risk stratification models

As we can validate the performance of the S-Learner using the observed outcomes, and due to its strong performance compared to the other methods in early experiments, we report the statistical analysis results of the S-Learner in this section and leave results for the X-Learner and the DR-Learner in **Appendix D.1.5**.

6.3.2 Experiments and Results

To study the mismatch between the objectives of current sepsis risk stratification approaches and the ultimate goal of allocating treatments to improve patient outcomes, we are interested in answering the following questions through our experiments:

- Is there heterogeneity in the effect of sepsis on mortality?
- How does the risk of developing sepsis correlate with the effect of sepsis on mortality?

Our final study population used for evaluation consisted of 7,282 ICU stays in the U-M cohort and 5,942 ICU stays in the BIDMC cohort. Information about the development

cohorts for each dataset in **Appendix D.1.4**. In the U-M evaluation cohort, 576 (7.9%) of ICU stays developed sepsis, and 574 (7.9%) admissions experienced in-hospital mortality. Within septic ICU stays, 126 (21.9%) experienced in-hospital mortality, while for non-septic ICU stays, 448 (6.7%) experienced in-hospital mortality. In the BIDMC evaluation cohort, 483 (8.1%) ICU stays developed sepsis, while 512 (8.6%) experienced in-hospital mortality. Within the septic group, 127 (26.3%) experienced in-hospital mortality, while 385 (7.1%) non-septic ICU stays experienced in-hospital mortality.

For the task of predicting the risk of developing sepsis, our learned models achieved an AUROC of an AUROC of 0.69 (95% CI: 0.67-0.71) in the U-M cohort and 0.74 (95% CI: 0.72-0.77) in the BIDMC cohort. For the task of predicting the risk of in-hospital mortality without sepsis and with sepsis, the S-Learner achieved AUROCs of 0.89 (95% CI: 0.87-0.90) and 0.79 (95% CI: 0.74-0.83) respectively in the U-M cohort and AUROCs of 0.87 (95% CI: 0.85-0.88) and 0.77 (95% CI: 0.73-0.82) respectively in the BIDMC cohort. The global null test shows that all models can accurately predict null treatment effects if they exist in the data, with the S-Learner performing the best (**Table 6.1**).

	U-M: Sepsis	**U-M: No Sepsis**	**BIDMC: Sepsis**	**BIDMC: No Sepsis**
S-Learner	0.00 (0.00-0.00	0.00 (0.00-0.00)	0.00 (0.00-0.00)	0.00 (0.00-0.00)
X-Learner	0.00 (0.00-0.00)	0.00 (0.00-0.00)	0.01 (0.01-0.01)	0.00 (0.00-0.00)
DR-Learner	0.01 (0.01-0.01)	0.00 (0.00-0.00)	0.02 (0.02-0.02)	0.00 (0.00-0.00)

Table 6.1: Global null results for all causal inference techniques across both datasets and when the model is trained on the septic admissions with random treatments and the non-septic admissions with random treatments.

To answer the first research question, we visualize the estimated effect of sepsis on mortality across all models and datasets (**Figure 6.6**). The S-Learner estimated a median effect of sepsis on mortality of 6.19 percentage points and 8.82 percentage points in the U-M cohort and the BIDMC cohort respectively. The entropy of the estimated effect of sepsis on mortality was 0.92 in the U-M cohort, and 0.87 in the BIDMC cohort. The X-Learner and DR-Learner showed similar heterogeneity (see **Appendix D.1.5**).

The Spearman's correlation between the estimated risk of sepsis and the estimated effect of sepsis on mortality showed a weakly positive relationship in both datasets (0.35 [95% CI: 0.33-0.37] and 0.31 [95% CI: 0.28-0.34]). Within quintiles of sepsis risk, there is large variability in the effect of sepsis on mortality within patient windows (**Figure** 6.7(a), **Figure** 6.7(b)). Many data points with a heightened sepsis risk might not experience severe consequences upon developing sepsis, while many at low sepsis risk could face significantly increased mortality risk if they were to develop sepsis. Among windows in the highest 20% risk of developing sepsis, sepsis was not estimated to substantially increase the risk of mor-

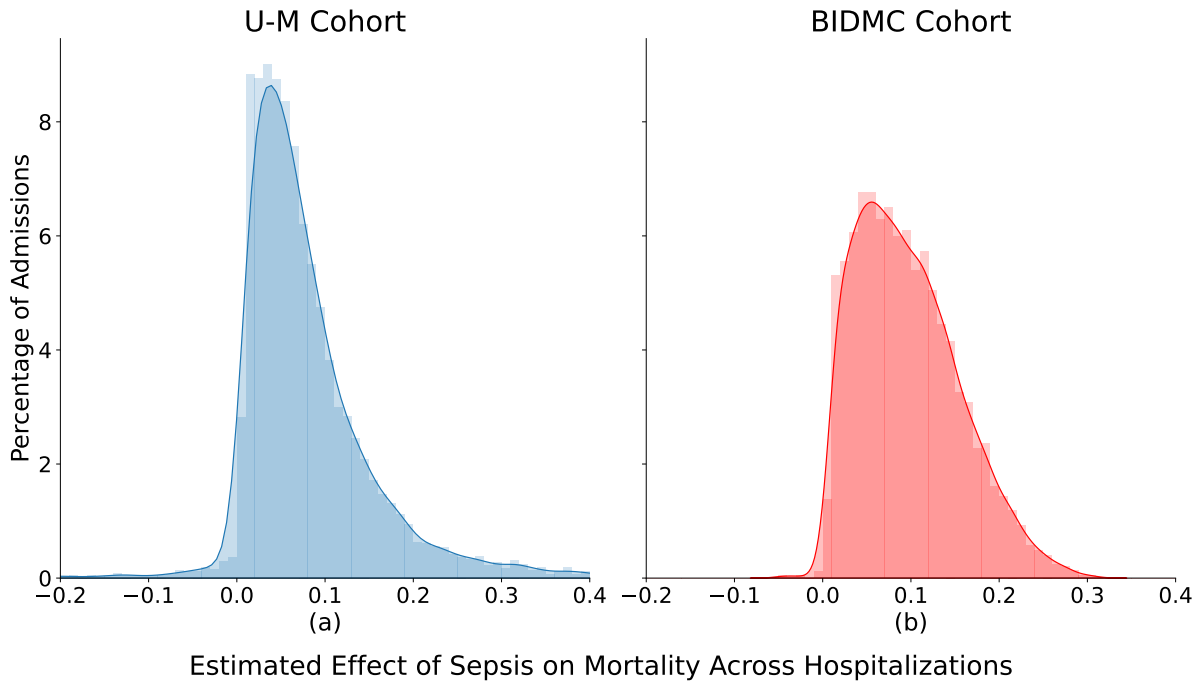

Figure 6.6: Estimated effect of sepsis on mortality across hospital admissions as estimated by the S-Learner. The average estimated effect is positive in both datasets. Moreover, there is substantial heterogeneity in the estimated effect of sepsis on mortality.

tality (i.e., < 5 percentage points) for 34.8% and 17.9% of windows in the U-M and BIDMC cohorts respectively (**Figure** 6.7(c), **Figure** 6.7(d)). Meanwhile, for the remaining 80% of windows, developing sepsis was estimated to have a substantial increase in mortality risk (i.e., > 20 percentage points) in over 7% of windows within each cohort. These overall findings hold for all other causal inference techniques, with most models showing a weak correlation between the risk of developing sepsis and the effect of sepsis on mortality (see **Appendix D.1.5**).

Overall, through our analysis, we found that: 1) there is substantial heterogeneity in the effect of sepsis on mortality and 2) those at a higher risk of sepsis are not necessarily more likely to experience mortality due to the development of sepsis. Our results bring into question the objective of current sepsis risk stratification tools and highlight the mismatch between how tools are currently built and the ultimate goal of augmenting decision-making towards improving downstream patient outcomes.

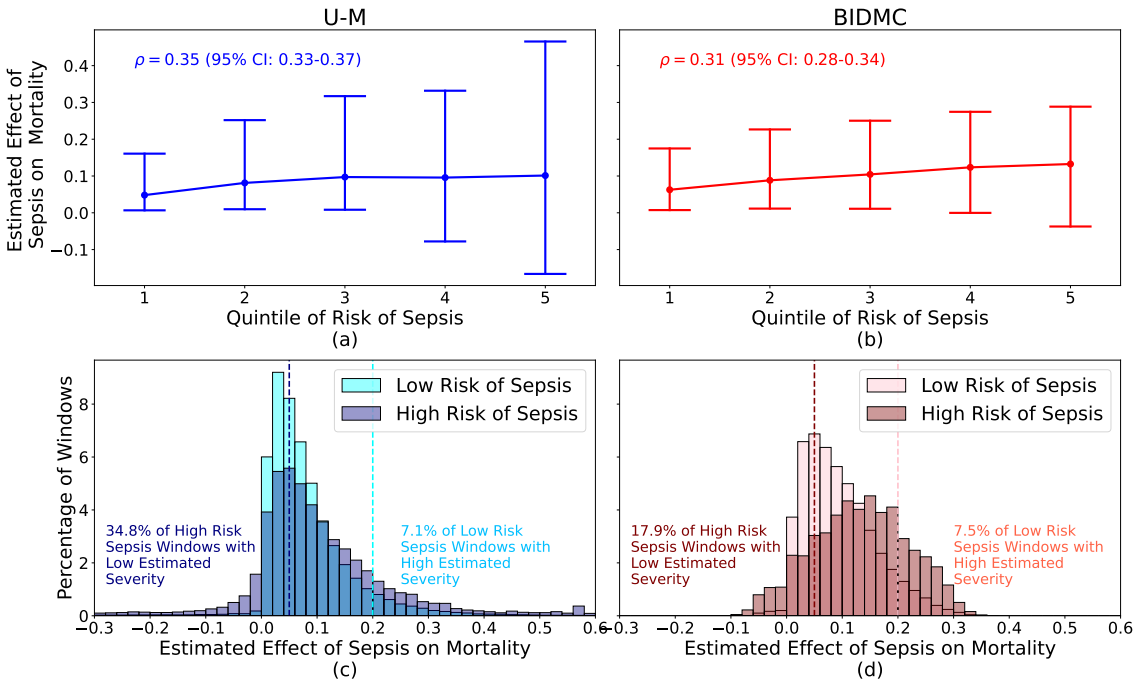

Figure 6.7: Relationship between the effect of sepsis on mortality, as estimated by the S-Learner, and the risk of developing sepsis. The estimated effect of sepsis on mortality is larger for windows within higher quintiles of risk of sepsis (top). Meanwhile, there are many high-risk sepsis windows that are still estimated to have a low effect of sepsis on mortality, yet many low-risk windows would be severely adversely affected by developing sepsis (bottom).

6.4 Discussion and Conclusion

Recent work has emphasized the need to accurately frame the development and evaluation of machine learning models to understand their potential for clinical impact [114]. In line with this idea, we perform a deep dive into existing risk stratification tools for sepsis. We focus on studying the mismatch between how current tools are built and evaluated and what is needed to augment clinical care and improve patient outcomes. First, we evaluated a commonly used sepsis risk prediction model, the ESM, with respect to when a clinician places an order for an indicator of treatment. We found that a majority of individuals who developed sepsis received some order for an indicator of sepsis treatment before they met the criteria for sepsis. Excluding predictions after treatment indicators, the model's performance was no better than random and performed significantly worse compared to using predictions up to meeting the sepsis criteria. This suggests that the ESM, a popular existing risk stratification tool, cannot help in identifying cases before clinical recognition. Next, we focused on understanding the limitations of the overall objective of existing risk stratification tools. Standard sepsis risk stratification approaches often focus on identifying patients at greatest risk of developing the disease. These approaches assume that prioritizing those most at risk of acquiring disease is optimal for reducing downstream mortality. We explored this assumption in the context of patient risk stratification for sepsis. Across both cohorts, the effect of developing sepsis on mortality was heterogeneous. Moreover, we consistently found that the risk of sepsis was only weakly correlated with the effect of sepsis on mortality. These findings held across both datasets and across causal inference techniques with only slight variability, pointing to an important limitation in standard sepsis risk stratification approaches. Overall, our findings highlight the importance of considering how risk stratification models will be used downstream when developing and evaluating ML models.

We began by focusing on an evaluation of existing risk stratification tools. We focused on the ESM due to its prevalence in healthcare systems across the US. The clinical utility and performance of the ESM has been a topic of recent interest across intuitions [20, 29, 125, 126]. In our study, we focus on measuring the performance of the ESM with respect to when a clinician places an order for a treatment indicator. In line with the type of evaluation proposed by Beaulieu-Jones et al., this helps to shed light on whether the ESM is simply relying on clinical intuition (i.e., looking over the shoulders of clinicians) or actually augmenting clinical knowledge in predicting the likelihood of sepsis [208, 19]. Building on this work, we present a new evaluation scheme for sepsis risk stratification that accounts for clinical intuition and evaluate a widely implemented sepsis risk model using this new evaluation scheme. Through our analysis, we found that clinicians ordered treatment before

sepsis criteria were met in a large majority of the sepsis population. Moreover, we found that the discriminative ability of the ESM during standard evaluation is attributable to predictions made after sepsis was clinically recognized and treatments were initiated, even when adjusting for the amount of clinical data available to the model. The poor performance of the ESM within this context helps to explain other recent work focused on evaluating the ESM, which have found that the ESM scores often do not cross the alert threshold for positive patients until after antibiotics are given or after lactate is measured [208, 20]. Moreover, the poor performance when evaluating using predictions before antibiotics helps explain findings in recent work by Burgin et al., who reported no improvement in the time to antibiotics for patients with sepsis when using the ESM [29]. These findings suggest that the ESM may not provide utility in guiding the timing of treatment before clinicians have already made that decision. Overall, this work shows the gap between existing evaluation schemes of sepsis risk stratification approaches and the goal of understanding the utility of a model for augmenting clinical decision-making.

We then studied the limitations of the objectives used to build existing sepsis risk stratification tools. When understanding how to allocate treatments in real clinical applications, recent work has considered estimating heterogeneous treatment effects [129, 85]. However, when understanding how to allocate a novel treatment that does not yet exist in the data, past work defaults to stratifying individuals by their likelihood of developing disease. These approaches assume that prioritizing those most at risk of developing the disease is optimal for reducing downstream mortality. We probed the validity of this assumption, finding that this assumption does not hold across two large clinical cohorts. The effect of sepsis on mortality was estimated to be heterogeneous. Moreover, the risk of sepsis was estimated to have a slight positive relationship with the effect of sepsis on mortality, indicating that those most likely to develop sepsis are more likely to experience mortality due to it. However, we found that this relationship was not strong, with correlations less than 0.5 across all causal inference methods and datasets. There are many windows of data with a high estimated risk of developing sepsis whose effect of sepsis on their mortality is quite low, and vice versa, there are many individuals who would not be classified as high risk of sepsis, whose mortality rate would greatly increase if they were to develop sepsis. Allocating interventions to the former rather than the latter could delay interventions to those who would most benefit, displaying the importance of considering downstream heterogeneous effects of disease when allocating new treatments and resources. These findings highlight the importance of considering the downstream effects of diseases on patient outcomes, towards the goal of improving the allocation and prioritization of treatments to improve individual health outcomes.

Though our analyses are focused on sepsis, our findings provide important implications

when developing and evaluating predictive models for other clinically relevant diseases. Predictive models are often built to estimate the likelihood of certain diseases in the hospital, such as *Clostridium difficile* and COVID-19 [144, 28]. These models follow the same principle of allocating treatments to those most at risk of developing the disease and evaluating with respect to the time of disease onset. However, the effect of these diseases on downstream complications, such as mortality, may be heterogeneous. Moreover, clinicians may recognize and treat signs of disease well before their deemed onset time. Our work complements past attempts to identify severe cases of disease and evaluate a model's ability to augment clinical intuition by performing an in-depth case study of existing sepsis risk stratification methodologies [24, 19]. The framework we consider throughout our studies can hence be used to study patient risk stratification in a multitude of different clinically relevant diseases.

Our study is not without limitations. First, we identified sepsis based on specific definitions and identified the ordering time of indicators for sepsis treatment according to this definition. However, sepsis definitions are still debated [172, 69, 95]. Next, we only assessed the ESM under the new evaluation scheme. Importantly, our goal was not to find the best sepsis risk model but rather to understand the limitations of current evaluation procedures for existing risk stratification tools. Moreover, when estimating the effect of sepsis on mortality, we assume a particular graphical model of the world. We stress that this model of the world is an oversimplification of the truly complex nature of sepsis and most diseases. However, we consider this a proof of concept for understanding the importance of modeling downstream heterogeneity in patient outcomes due to disease. Next, as we are estimating causal effects from observational data, our causal inference techniques rely on stringent assumptions that are untestable in the data. We cannot accurately measure whether we have sufficient overlap between the sepsis and no sepsis populations and whether we have measured all relevant confounders that may affect the likelihood of developing sepsis and the effect of sepsis on mortality. Violations of these assumptions may result in biased treatment effect estimates. Finally, due to the lack of ground-truth treatment effects, we are unable to accurately validate the learned effect of sepsis on mortality, despite the use of a proxy evaluation. To overcome this, we follow past work and ensure that our key takeaways hold across a multitude of different causal inference techniques [210, 112].

Overall, our findings have significant implications for the development and evaluation of clinically useful models for sepsis prediction. When building predictive models for identifying individuals for whom to prioritize treatment, researchers often ignore the potentially heterogeneous effects that the acquisition of the disease may have on downstream patient outcomes. Moreover, researchers often evaluate these models with respect to disease onset, rather than with respect to clinical intuition. Our findings emphasize the limitations

of these approaches and highlight the importance of considering the mismatches between existing tools and how they may be used in clinical settings.

CHAPTER 7

Conclusion

This dissertation focused on the mismatch between what ML models are optimized and evaluated for and what is needed in certain clinical contexts for risk prediction and resource allocation. The lack of adoption of existing techniques is likely due to a variety of contributing factors, including a gap between ML objectives and clinical needs. ML models are often optimized and evaluated according to common benchmark tasks. However, the specific needs of a healthcare worker in certain contexts might differ from what an ML model is optimized for. For example, past work often focuses on accurate treatment effect estimation. However, in many scenarios, clinicians simply need an accurate ranking of individuals ordered by their benefit from treatment. This mismatch can result in sub-optimal ML models and slow adoption of these models. Our primary thesis centers around the idea that specific clinical needs can and should inform training and evaluation of ML models for greater clinical impact.

In this dissertation, we explored issues slowing the adoption of ML algorithms in clinical practice and presented new approaches towards bridging the gap between research and practice. We focused on the problems of risk prediction and resource allocation due to their potential for improving decision-making and impacting clinical care and provided evidence that such mismatches persist in existing tools and research. Our work shows that researchers should continue studying the mismatch between models and clinical needs across other tasks at the intersection of machine learning and healthcare towards more adoption of meaningful ML models to augment clinical workflows and improve patient outcomes. Overall, our work builds on research spanning several fields, including survival analysis and causal effect estimation. We summarize our contributions and place them in the context of the broader literature below.

First, when building survival analysis models, good calibration is an essential aspect for personalized and individualized decision-making. In particular, calibration can be vital to help patients and healthcare professionals make life decisions in anticipation of some health event. However, as discussed in **Chapter 3**, past work has focused on training and evaluating for discriminative performance, achieving state-of-the-art results by utilizing deep learning

models without the use of distributional assumptions [116, 156]. However, discriminative performance alone does not ensure good calibration. To address this gap, **we presented a framework for training and evaluating deep survival models that focuses on both calibration and discriminative performance.** We provided a theoretically sound approach for training deep survival models that, when applied in the context of a state-of-the-art neural network architecture, led to significant gains in the trade-off between calibration and discriminative performance across two publicly available clinical datasets. This work cautions against overfitting to one particular metric when training deep survival models and encourages model developers to adopt a more comprehensive evaluation that better aligns with potential clinical utility. Overall, this work represents a step towards the use of survival analysis models to augment clinical decision-making for risk prediction.

Next, we move from risk prediction to intervention allocation and study the mismatch between techniques in causal inference and the needs of practitioners. When considering intervention allocation in resource-constrained settings, practitioners often wish to understand who would benefit most from a particular treatment. Hence, past work focuses on accurately estimating CATEs from observational data to help rank individuals. However, in many scenarios, practitioners simply require a ranking of individuals by most benefit. **In Chapter 4, we studied the objective mismatch between accurate CATE estimates and an accurate ranking of individuals** when the goal is maximizing benefit across all treatment thresholds. We showed that accurate CATE estimates are a sufficient but not necessary condition for optimal expected benefit and that better CATE accuracy does not necessarily correspond to a better ranking. We presented a novel approach for directly optimizing for ranking and, through an empirical case-study, we showed the efficacy of optimizing for expected benefit for treatment allocation at low sample sizes across two synthetic datasets. This work is an important step for bridging the theory and practice of resource allocation techniques and highlights the potential for sub-optimality of current ML approaches when not considering clinical context.

Third, in situations where accurate CATEs are necessary for decision-making, there exists a multitude of approaches for estimating accurate causal effects. All methods aim to overcome issues due to confounding, which if left unaddressed, can lead to biased and inaccurate CATE estimates. As shown in **Chapter 5**, popular techniques can often be classified into three different categories, each with its theoretical strengths and guarantees. Theoretical results often assume infinite data, and hence, how this theory translates into a variety of different settings in practice has been under-explored. In Chapter 5, **we presented an extensive empirical exploration of popular CATE techniques** in the context of deep learning to better understand the mismatch between theoretical results and empiri-

cal performance in practical settings. We found that popular approaches that adjust for the propensity score, including those that incorporate estimates of the potential outcomes as well, were unable to consistently outperform techniques that simply rely on only estimates of the outcomes. This work shows the importance of extensive validation of theoretical results in realistic settings to understand how theory may translate to practice.

Finally, to close the gap between our contributions in the field of ML and the ultimate goal of improving clinical practice, we carefully studied the problem of sepsis risk stratification, exploring real-world mismatches between existing risk stratification tools and clinical care needs. Sepsis risk stratification is a well-studied problem in the field of ML [208, 77, 44, 207, 106]; however, there exist important gaps that preclude impact in clinical settings. **In Chapter 6, we studied these gaps in both model development and evaluation for sepsis risk stratification**. We first found that when evaluating sepsis risk stratification models with respect to clinical recognition rather than the time of meeting sepsis criteria, a widely used technique fails to perform significantly better than random. Second, we found that sepsis risk stratification models that only focus on the likelihood of getting sepsis may be sub-optimal, as there exists heterogeneity in the effect of sepsis on mortality and this heterogeneity is not strongly correlated with the likelihood of developing sepsis. Our work highlights how the needs of clinicians can and should inform ML model development in healthcare settings.

There are several areas discussed throughout this dissertation that could be interesting for future work. Here, we outline four possibilities.

First, the fundamental problem in causal inference eliminates the ability to use ground truth treatment effects during training and evaluation. In many fields such as estimating the individualized effects of antibiotics, there exist techniques to accurately estimate the counterfactual of what would have happened if an individual was given a different treatment [23, 102]. However, obtaining counterfactual annotations may be time-consuming and expensive, and annotations may only be available for a potentially biased sub-population of the data. Hence, there exists an opportunity to: 1) learn how to collect missing counterfactuals while balancing cost and potential increase in accuracy, and 2) learn how to best leverage both a cohort of individuals with ground-truth treatment effects and those with only observed outcomes. The former can build upon ideas from research in active learning for CATE estimation, which focuses on learning to defer and labeling observed outcomes [88, 89]. The latter could build upon ideas from both semi-supervised learning and research focused on learning causal effects leveraging both randomized controlled trials and observational data [97, 36, 76]. Combined, such a pipeline could dramatically improve the ability to learn CATEs in many clinical settings where ground-truth treatment effects can be collected.

Second, a limitation of much work in causal effect estimation is the need for assuming no hidden confounders, overlap, and consistency. As these assumptions remain untestable, it is impossible to understand whether these assumptions hold when utilizing causal inference techniques in downstream applications. Though there has been some work in building theory and models to overcome violations of these assumptions, they often require stringent assumptions of their own [123, 161, 98, 132]. While identifiable CATEs imply identifiable rankings, violations of these assumptions may not always render learning optimal rankings from the data impossible, even if CATEs are unidentifiable. Past work has studied relaxing unconfoundedness and developed the rank-preserving assumption (RPA), showing that estimates of CATES from data with unobserved confounding can still ensure optimal rankings if they are rank-preserving [51]. Under these conditions, ranking remains identifiable, and such biases may even simplify the ranking problem [51]. We can extend this idea to violations of consistency, where noise in observed outcomes may result in unidentifiable CATEs, but under rank-preserving assumptions, may still result in accurate rankings. This approach is similar to the boundary-consistent noise models used in traditional classification problems and opens an interesting new area for research [131].

Third, another important limitation of our work in causal effect estimation is the inability to validate models using real-world data. Though some work exists towards overcoming this issue, they often require the use of proxy variables or data that must be estimated from the data and may be inaccurate [211, 148]. Errors in the problem set-up and evaluation may easily result in inaccurate takeaways, with the potential to lead to harm when applied to real data. As the goal of these models is to impact clinical care, there exists an opportunity to incorporate clinical experts into the validation scheme of causal inference techniques. This may be through manual inspection or through understanding how decision-making is impacted when augmented using different techniques. Such a validation can also help understand whether differences between model performance in synthetic settings, as measured by mean squared error, result in meaningful differences when augmenting clinical decision-making. Hence, there exists an opportunity to formalize such a pipeline towards a more standardized evaluation of causal inference techniques in real medical data.

Finally, our work in survival analysis is limited to working with retrospective data. Accordingly, we were unable to validate whether the proposed survival models resulted in a meaningful impact for improved clinical care. Though focusing on calibration is an important step towards this goal, future work should consider how survival models should be integrated into the clinical workflow and what other aspects of learned survival models could be improved for use by clinical experts.

The main contributions of this dissertation are 1) a holistic framework for training and

evaluation of deep survival models for both calibration and discriminative performance, 2) a theoretical and empirical case study of the efficacy of optimizing directly for maximizing benefit for treatment allocation, 3) an extensive comparison of popular CATE estimation techniques across a variety of practical settings, and 4) a demonstration of the gap between the development and validation of sepsis risk stratification models and the goal of augmenting clinical users and improving patient outcomes. Going forward, we expect problems studied in this dissertation to help take an important step towards ML having real clinical impact when augmenting decision-makers for risk prediction and intervention allocation.

APPENDIX A

Appendix for Calibrated Deep Survival Analysis

A.1 Deep Survival Analysis Architectures

Recently, many have applied neural networks to data with censored individuals for survival analysis [124, 104, 155, 5]. However, many of these models rely on assumptions about the distributional form of the time-to-event data, such as the proportional hazards assumption [41, 201]. These assumptions may not generalize to new data. Accordingly, we focus our analysis on deep survival analysis architectures that achieve state-of-the-art discriminative results without explicitly relying on any distributional assumptions. Despite reported gains in discriminative performance, to date, these models have not been evaluated in terms of calibration.

DeepHit was one of the first fully distribution-free methods for survival analysis [116]. DeepHit corresponds to a feed-forward neural network architecture that takes as input an individual's covariates $\mathbf{x}_i$, and outputs a probability distribution $\hat{\mathbf{y}}_i \in [0,1]^\tau$, where $\hat{y}_{i,t}$ corresponds to the estimated $\hat{P}(Z = t|\mathbf{x}_i)$. The CIF at time t can then be estimated as $\hat{F}(t|\mathbf{x}_i) = \sum_{j=1}^{t} \hat{y}_{i,j}$. The final layer of DeepHit is a softmax output layer requiring $\hat{F}(\tau|\mathbf{x}_i) = 1$. This formulation assumes that by the end of the time horizon τ, every individual will have had the event. Hence, this formulation will incorrectly estimate the true underlying survival process for individuals who survive beyond time τ. Moreover, as DeepHit outputs a fixed-sized vector, it can not be used to forecast survival curves past the specified time horizon τ.

DRSA, or deep recurrent survival analysis, alleviates this structural issue of DeepHit while taking advantage of the sequential patterns present in survival analysis [156]. DRSA uses a long short-term memory (LSTM) network that takes as input at timestep t, a concatenation of an individual's covariates $\mathbf{x}_i$ and t [80]. The output of the LSTM at time t is passed into a fully connected layer with a sigmoid activation function that

outputs $\hat{\lambda}(t|\mathbf{x}_i)$. Accordingly, we can estimate the survival probability at timestep t as $\hat{S}(t|\mathbf{x}_i) = \prod_{j:j\leq t}(1 - \hat{\lambda}(j|\mathbf{x}_i))$, and the probability of the event occurring at timestep t as $\hat{P}(Z = t|\mathbf{x}_i) = \hat{\lambda}(t|\mathbf{x}_i)\prod_{j<t}(1 - \hat{\lambda}(j|\mathbf{x}_i))$. Since DRSA does not make assumptions about the probability of survival at the end of the horizon while still allowing for variable-length forecasting of survival curves, we build on this architecture in our proposed approach.

A.2 Full Proof that $\mathcal{L}_{RPS}$ Elicits Calibrated Survival Curves

Claim. *Training deep survival models using $\mathcal{L}_{RPS}$ will result in well-calibrated estimates of survival.*

Proof. Consider n individuals with identical or near-identical covariates with observed event times $\{z_i\}_{i=1}^n$. Define the counting-based Kaplain-Meier estimate for these individuals at time t as $KM_t^n = \frac{1}{n}\sum_{i=1}^n 1_{t<z_i}$, where $\lim_{n\to\infty} KM_t^n$ is the underlying survival probability at time t for these n individuals.

A survival model will estimate one survival probability for these n individuals at time t. Define this value as $\hat{p}_t$. A well-calibrated survival model will output a $\hat{p}_t$ that closely aligns with the underlying survival probability $\lim_{n\to\infty} KM_t^n$. Consider the optimization problem of finding $\hat{p}_t$ which will minimize $\mathcal{L}_{RPS}$. This problem can formally be set-up as $\arg\min_{\hat{p}_t}\sum_{i=1}^n(\hat{p}_t - 1_{t<z_i})^2$.

First, this optimization problem is strictly convex and has a unique minimum, as the second derivative is positive everywhere. such that any minimizer must be the unique minimizer to this loss function. In order to do so, consider taking the second derivative of the objective function with respect to $\hat{p}_t$.

$$\frac{\partial^2}{\partial \hat{p}_t^2}\left(\sum_{i=1}^n(\hat{p}_t - 1_{t<z_i})^2\right) =$$
$$\frac{\partial}{\partial \hat{p}_t}\left(2\hat{p}_t - \frac{2}{n}\sum_{i=1}^n 1_{t<z_i}\right) =$$
$$2 \geq 0$$

To find the value of $\hat{p}_t$ that minimizes this objective function ($\hat{p}_t^*$), we set the derivative equal to zero.

$$\frac{\partial}{\partial \hat{p}_t^*}\left(\sum_{i=1}^{n}(\hat{p}_t^* - 1_{t<z_i})^2\right) = 0$$

$$2\hat{p}_t^* - \frac{2}{n}\sum_{i=1}^{n} 1_{t<z_i} = 0$$

$$\hat{p}_t^* = \frac{1}{n}\sum_{i=1}^{n} 1_{t<z_i}$$

The unique estimated survival probability that minimizes the objective function is equivalent to the average survival status for all n individuals at time t. This unique minimum is equal to KM_t^n which, as n gets large, is equal to the true underlying survival probability for these individuals at time t. Hence, training a survival model to minimize $\mathcal{L}_{RPS}$ will result in estimated survival probabilities that align well with the true survival probabilities. □

A.3 Censored DDC

In the case of censored individuals, we only know that prior to censoring the event did not occur. Following the probability integral transform argument used to justify DDC, for a well-calibrated model, we would expect half of the individuals to have the event after reaching an estimated survival probability of 50%. If *more* than half the individuals are censored after reaching an estimated survival probability of 50%, then we can conclude that the model is *not* well-calibrated. However, if *less* than half of the individuals are censored after reaching an estimated survival probability of 50%, we cannot conclude anything with respect to model calibration (the event may take place at any time after censoring). Given these limitations, without strong assumptions on the event time distribution for censored individuals, one cannot make meaningful conclusions regarding the calibration of a model for censored individuals. To this end, while we measure discriminative performance across both uncensored and censored individuals, we focus our evaluation of calibration on uncensored individuals.

A.4 Trade-Off Between Discriminative Performance and Calibration

To display the trade-off between discriminative performance and calibration, we simulate 1,000 covariates and corresponding sampled event times through the following scheme:

$$\mathbf{X} = (\mathbf{X}^a, \mathbf{X}^b)^T \in \mathbb{R}^{1,000\times 20}$$
$$\mathbf{X}^a = (\mathbf{X}^a_1, \mathbf{X}^a_2) \in \mathbb{R}^{500\times 20}$$
$$\mathbf{X}^b = (\mathbf{X}^b_1, \mathbf{X}^b_2) \in \mathbb{R}^{500\times 20}$$
$$\mathbf{X}^a_1, \mathbf{X}^b_1 \sim U(0, 10)^{10}$$
$$\mathbf{X}^a_2 \sim U(10, 20)^{10}$$
$$\mathbf{X}^b_2 \sim U(5, 15)^{10}$$
$$z_i \sim LN(.5(\mathbf{1}^T\mathbf{x}_i^{1:10})^2 + 2(\mathbf{1}^T\mathbf{x}_i^{11:20})^2, 0.5)$$

Note that U and LN denote a uniform and a log-normal distribution respectively. We consider τ (the time horizon) to be the 50th percentile of sampled event times, in order to right-censor half of the individuals. Finally, we place all time to events into one of 100 equally spaced time bins.

Given this simulation, we calculate the C-index value for the ground-truth log-normal survival curves. The average C-index of the ground-truth survival curves in these finite samples across 1000 replications of the simulation is .760 (95% Confidence Interval: (.742, .778)). This is due to examples such as the one displayed in the **Figure A.1**. Though an individual can experience an event early, it is not necessarily true that their true survival probability is low. These situations result in incorrect rankings among different individuals, which contributes negatively towards the C-index value.

Importantly, we note that this is due to the single sample definition of discrimination. For example, for a particular observed outcome distribution, it is possible to achieve perfect discrimination (as measured by the C-index) by estimating Heaviside distributions that drop to 0 at the observed event times. However, these distributions do not take into account the stochasticity that likely exists in the survival process. Due to this stochasticity, it is unlikely for the underlying survival curves to provide perfect discriminative performance (i.e. a C-index of 1) with respect to the observed outcomes, showing an important trade-off that is necessary to consider when evaluating survival models.

A.5 Additional Experimental Set-Up Details

Dataset Details We consider two public clinical datasets: the Northern Alberta Cancer Dataset and the CLINIC dataset. For each dataset, we use the same 60/20/20% train/validation/test split across model initializations in order to train and evaluate our models. We stratify our random splits in order to ensure a roughly equal proportion of censored individuals in each split. We normalize all covariates by the mean and standard

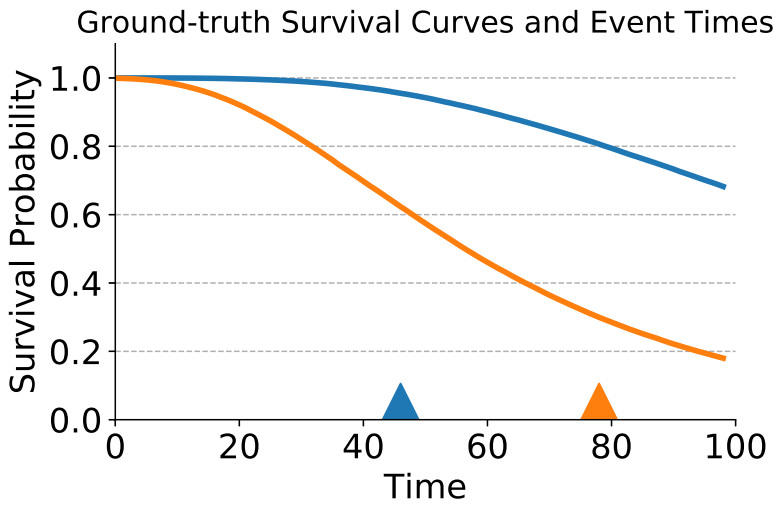

Figure A.1: An example pair of ground-truth survival curves for 2 individuals from a simulated stochastic process. Triangles denote the observed event times. As the blue individual experienced the event at a high survival probability, they will consistently be ranked incorrectly when compared to other individuals who have a lower survival probability but experience the event later (*e.g.*, the orange individual). These examples will contribute negatively to the C-index evaluation, despite good calibration.

deviation of each feature in the training set.

Additional Baselines. For completeness, we report the results for two additional baseline methods. Namely, we train two variants of the feed-forward DeepHit model. First, we train the DeepHit architecture with the loss as it was originally proposed ($\mathcal{L}_{log} + \lambda\mathcal{L}_{kernel}$). To examine the importance of $\mathcal{L}_{kernel}$ in DeepHit and examine the performance of $\mathcal{L}_{log}$ alone, we also consider evaluating the performance of DeepHit without the kernel loss ($\lambda = 0$).

Additional Training and Hyperparamter Details. All DRSA models had the same architecture: a one-layer LSTM with hidden size 100 and a single feed-forward layer with a sigmoid activation on the output for each time-step. For DeepHit, we followed the same architecture proposed in the original paper. We considered learning rates of *1e-3* and *1e-4*, but preliminary results found no comparable difference in performance on the held-out validation set, so we continued using a learning rate of *1e-3*. In order to tune the σ hyperparameter for the $\mathcal{L}_{kernel}$ loss function, we considered σ values from 0.1 to 10. σ was then chosen based on performance on the held-out validation set on the NACD dataset. This optimal σ value ($\sigma = .8$) was used for both the NACD dataset and the CLINIC dataset in order to test the generalizability of the relationship between $\mathcal{L}_{RPS}$ and $\mathcal{L}_{kernel}$ in the composite loss. Other hyperparameters, were chosen based on performance on the held-out validation set as well. Due to the right-skewed time-to-event distribution which can cause $\mathcal{L}_{RPS}$ to ignore earlier

Table A.1: Discriminative (C-index) and calibration performance (DDC, D-Calibration, Averaged Brier Score), as well as the trade-off between the two (total score) for the NACD and CLINIC datasets (mean ± standard deviation across random initializations, number of times passing the statistical test for D-Calibration). Lower DDC and Brier score values indicate better performance, while higher values of C-index, D-Calibration, and total score indicate better performance. The proposed training approach consistently leads to improvements in calibration, without sacrificing discriminative performance or Brier score. An * indicates results that are statistically significant over all baselines using a paired t-test ($p < .05$).

Model	**NACD**				
	C-index ↑	**DDC ↓**	**D-Calibration ↑**	**$\overline{\text{Brier}}$ ↓**	**Total Score ↑**
Ren et al. 2019	.748 ± .002	.025 ± .012	1	.101 ± .002	.846 ± .004
MTLR	.750 ± .000	.062 ± .000	0	.101 ± .000	.834 ± .000
DeepHit ($\mathcal{L}_{log}$)	.751 ± .002	.083 ± .005	0	.102 ± .000	.826 ± .003
DeepHit ($\mathcal{L}_{log} + \lambda\mathcal{L}_{kernel}$)	.748 ± .004	.020 ± .005	0	.107 ± .001	.849 ± .003
Proposed - $\mathcal{L}_{RPS}$	.741 ± .008	.305 ± .089	0	.207 ± .034	.715 ± .050
Proposed - $\mathcal{L}_{kernel}$	.742 ± .003	.012 ± .002	3	.101 ± .003	.847 ± .001
Proposed Method	.742 ± .006	**.007 ± .003***	**5**	.104 ± .002	.850 ± .003
Model	**CLINIC**				
	C-index ↑	**DDC ↓**	**D-Calibration ↑**	**$\overline{\text{Brier}}$ ↓**	**Total Score ↑**
Ren et al. 2019	.616 ± .003	.138 ± .002	0	.107 ± .000	.719 ± .003
MTLR	.608 ± .000	.168 ± .000	0	.106 ± .000	.702 ± .000
DeepHit ($\mathcal{L}_{log}$)	.616 ± .003	.133 ± .004	0	.103 ± .000	.720 ± .002
DeepHit ($\mathcal{L}_{log} + \lambda\mathcal{L}_{kernel}$)	.624 ± .001	.063 ± .007	0	.106 ± .001	.749 ± .002
Proposed - $\mathcal{L}_{RPS}$	.628 ± .003	.241 ± .022	0	.153 ± .002	.687 ± .011
Proposed - $\mathcal{L}_{kernel}$	.615 ± .005	.097 ± .006	0	.110 ± .001	.731 ± .005
Proposed Method	.627 ± .001	.056 ± .011	0	.106 ± .001	.753 ± .004

time-points before time-to-events, we up-weighted these earlier time-points to provide equal supervision across the horizon. In order to tune the regularization constants of MTLR, which control the amount of smoothing for the model, we used the cross-validation scheme built into the MTLR R package.

A.6 Additional Results

The proposed method continues to consistently outperform all baselines with respect to DDC and D-calibration while maintaining comparable C-index and average Brier score values (**Table A.1**). Compared to DRSA and DeepHit with $\lambda = 0$, the proposed method results in a statistically significant improvement in calibration across both tasks (NACD DDC: .025 and .083 vs. .007, CLINIC DDC: .138 and .133 vs .056). This improvement, however, is accompanied by a small decrease in C-index in the NACD dataset. Moreover, training using $\mathcal{L}_{RPS}$ alone results in better calibration than both DRSA and DeepHit trained using only

$\mathcal{L}_{log}$ (NACD DDC: .025 and .083 vs .012, CLINIC DDC: .138 and .133 vs .097), with minimal drops in discriminative performance. These empirical results support the original hypothesis that training using $\mathcal{L}_{RPS}$ should result in survival models that better balance discriminative performance and calibration.

DeepHit that includes training with $\mathcal{L}_{kernel}$ consistently results in better calibration compared to DeepHit without this loss function (DeepHit ($\lambda = 0$)). This supports the hypothesis that $\mathcal{L}_{kernel}$ can act as a scaling mechanism to calibrate survival estimates without sacrificing discriminative performance. Despite this increased performance, our proposed approach still achieves better calibration performance (NACD DDC: .020 vs .007, CLINIC DDC: .063 vs .057), while also maintaining a better trade-off between calibration and discriminative performance, as shown through the total score.

Overall, these results continue to support our original hypothesis regarding the efficacy of the training scheme. We show that training using $\mathcal{L}_{RPS}$ outperforms models that solely train using $\mathcal{L}_{log}$, while including the kernel loss function can consistently improve calibration performance with respect to DDC and D-Calibration. Finally, the best performance consistently comes from our proposed method, the combination of $\mathcal{L}_{RPS}$ and $\mathcal{L}_{kernel}$.

APPENDIX B

Appendix for Learning to Rank for Treatment Allocation

B.1 Related Work

CATE Estimation. In recent years, there has been increased interest in estimating the heterogeneous effects of treatments from confounded observational data [214]. A majority of past works have proposed solutions for overcoming the issue of confounding. Past work has considered learning balanced representations [182, 91, 92, 75], reweighting using propensity scores [74, 75, 12, 118], and using doubly robust proxies [105] across a wide variety of machine learning architectures, namely neural networks [182] and random forests [200]. However, these works tend to optimize for and evaluate the performance of techniques for their ability to accurately estimate CATEs. However, in finite samples when these models are not perfect, how performance, as measured by accuracy, translates to maximizing benefit has not been well-explored. Finally, past work has considered evaluating treatment effects under different resource constraints [173]. However, this work has focused on estimating the ATE under different potential treatment strategies, while we focus on the goal of understanding who to treat across different potential treatment thresholds.

Causal Decision Making. There has been recent interest in how causal inference techniques may translate to downstream decision-making. Recent work has studied when causal effect estimation may be insufficient when the goal is to identify whom to treat and framed a new problem of causal classification for identifying treatment responders [52, 14, 96]. This path represents a step towards bridging the gap between theory and practice for causal inference. In our work, we extend this idea even further beyond a binary classification problem and study the problem of optimal ranking policies without the need for an a priori threshold to label individuals as responders or non-responders [212]. As these thresholds for defining responders vs. non-responders may vary depending on the application, and may change many times for the same application, it remains essential to build models agnostic to

a particular threshold. Recent work has studied how confounded data may affect the task of ranking causal effects [51]. In our work, we continued with the no hidden confounders assumption and focused on building a technique for optimal ranking for maximizing benefit.

Uplift Modeling. Uplift modeling is the field of work closely related to our setting. Uplift modeling focuses on directly targeting interventions and measuring incremental gain as individuals become intervened upon [171, 21]. Uplift modeling is a common method used particularly in business and marketing problems [171, 212]. One approach towards uplift modeling is to estimate pointwise effects of interventions on an individual basis, similar to CATE estimation [70, 139]. A secondary approach is to optimize for cumulative gain across intervention thresholds, similar to our goal [222, 46]. However, uplift modeling uses data obtained from a randomized controlled trial, and hence, methods for optimizing for cumulative gain are not built to handle confounded data. For example, contextual treatment selection is built under the assumption of randomness, and build approximations to optimize for under this assumption [222]. In our work, we extend ideas from uplift modeling to directly optimize for optimal rankings for maximum benefit when learning from observational data. Moreover, we study optimizing for optimal rankings for maximum benefit across all potential treatment thresholds as defined by the AUTOC in the context of resource constraints where treatment may benefit everyone, a problem not studied in past work. Perhaps most similar to our work is recent work by Zhou et al [223]. Though they also consider the problem of ranking, their work differs in several ways. First, Zhou et al. focus on a setting in which randomized controlled trials are available. However, we focus on expanding the idea of ranking for accurate treatment allocation based on maximizing expected benefit to settings with only observational data (e.g., much of healthcare). Though techniques like inverse weighting using the propensity score can be used in observational data settings, it is not immediately obvious how one should adapt the approach proposed by Zhou et al. to the observational setting. Second, we demonstrate the benefit of directly optimizing for treatment allocation as defined by maximizing expected benefit compared to accurate CATE estimates. We focus on a theoretical and empirical exploration of the disconnect between these two problem set-ups. Meanwhile, the loss function in Zhou et al. relies on converging to an unbiased CATE estimate to correctly order individuals, and hence, does not directly optimize for treatment allocation. We present a case study to show how and when direct optimization may be of most benefit through our empirical results.

Learning to Rank (LtR). LtR methods focus on learning optimal rankings, particularly for search relevancy problems [32]. Pointwise methods, which estimate the exact relevancy of a document for a query, remain analogous to a majority of past work in CATE estimation. However, past literature in the field of LtR has also focused on pairwise techniques, which

focus on learning optimal ordering for pairs of inputs, and listwise techniques, which aim to directly optimize a list of inputs towards a measure of downstream measure of performance, either through direct optimization of using proxy loss functions [82, 32, 209, 186]. A common measure of performance studied thoroughly is the normalized discounted cumulative gain (NDCG), focused on recommending the most relevant items to a query first [87, 203]. The NDCG is a commonly accepted metric in the LtR field but does not have a meaningful interpretation for our setting in measuring the expected benefit from treatment across all thresholds u. Meanwhile, AUTOC measures both the ranking of examples as well as the cumulative treatment effect across any policy. Listwise learning to rank techniques have recently been studied for the related field of uplift modeling. However, these methods often assume binary outcomes from randomized controlled trials, two limitations unsuitable for our general application [46, 21]. In our work, we take inspiration from the field of listwise techniques built for optimizing NDCG and study how to extend these methods towards the problem of maximizing benefit for resource allocation, as measured by AUTOC, when learning from observational data.

B.2 Additional Proofs

(Restated) Proposition 1. *There exists a function $f \in \mathcal{F}$ such that $AUTOC_S(f) = AUTOC_S(f^*)$, yet $\mathcal{L}_S^M(f) > 0$.*

Proof. Define $f(\mathbf{x}_i) = f^*(\mathbf{x}_i) + \frac{\gamma_i}{3}$. Note that for this f, we have that $AUTOC_S(f) = AUTOC_S(f^*)$, yet:

$$\begin{aligned}\mathcal{L}_S^M(f) &= \frac{1}{n}\sum_i (f(\mathbf{x}_i) - \tau_i)^2 \\ &= \frac{1}{n}\sum_i (f^*(\mathbf{x}_i) - \frac{\gamma_i}{3} - \tau_i)^2 \\ &= \frac{1}{n}\sum_i (\frac{\gamma_i}{3})^2 > 0\end{aligned}$$

□

(Restated) Proposition 2. *For any model f such that $\mathcal{L}_S^M(f) > 0$, there exists a model g such that $\mathcal{L}_S^M(f) < \mathcal{L}_S^M(g)$ and $AUTOC_S(g) > AUTOC_S(f)$*

Proof. We may build a model g that achieves perfect ranking, but arbitrarily poor $\mathcal{L}_S^M(g) = C$ as follows: 1) Define α such that $\sum_{i=1}^n \alpha^2 = C$, and 2) $\forall \mathbf{x}_i, g(\mathbf{x}_i) = f^*(\mathbf{x}_i) + \alpha$. Note that

$AUTOC(g) = AUTOC(g^*)$, yet:

$$\begin{aligned}\mathcal{L}_S^M(g) &= \frac{1}{n}\sum_i (f(\mathbf{x}_i) - \tau_i)^2 \\ &= \frac{1}{n}\sum_i (f^*(\mathbf{x}_i) - \alpha - \tau_i)^2 \\ &= \frac{1}{n}\sum_i (\alpha)^2 = C\end{aligned}$$

Setting C to be larger than $\mathcal{L}_S^M(f)$ leads to the desired result.

□

B.3 Methods

In **Algorithm 1**, we describe the proposed splitting procedure at any decision node M. We choose features and corresponding values to split on that result in trees that maximize the proxy of the AUTOC when considering all samples in the data.

Algorithm 1 Calculating Split Value to Maximize AUTOC

Input: S: Complete dataset; S_M, T^M : Current dataset and tree at decision node M
Output: Feature k and value v to split data for maximizing AUTOC

Calculate **best value** as $\widetilde{AUTOC}_S(T^M)$ by traversing sample S through current tree T^M
for k,v in S_M that result in valid partitions **do**
  Build $T_{k,v}^M$ by splitting current node M by feature k and value v
  Calculate **proposed value** as $\widetilde{AUTOC}_S(T_{k,v}^M)$ by traversing sample S through $T_{k,v}^M$
  if proposed value improves over **best value then**
    Update **best value** to **proposed value**
    Update **best k,v** to be **proposed k,v**
return best k and v if they exist

Table B.1: Hyperparameters and their corresponding search ranges.

Hyperparameter	**Hyperparameter Search Range**
Number of Trees	100, 200, 500, 1000
Data Subsample Proportion	0.1, 0.2, 0.45, 1
Maximum Depth	3, 5, 10, 20, ∞
Minimum Examples in Node to Split	2, 5, 10, 20, 40
Minimum Examples in Leaf	1, 2, 5, 10, 20
Improvement Threshold	0, None

B.4 Experimental Set-Up

Model Training. Our proposed and baseline methodologies consist of two steps: 1) Build doubly robust proxies for training, and 2) Train a random forest algorithm using a certain split procedure using the doubly robust proxies as imputed CATEs. The doubly robust proxies are shared between both methods, ensuring that any difference between the two techniques is not due to the accuracy of these proxies. To give all methods the best opportunity to learn, we use cross-fitting with decision trees to estimate the potential outcomes and use accurate propensity scores to build the doubly robust proxy. For the second step, we train all methods using the same underlying random forest architecture, while only varying the split procedure. When building each decision tree within the random forest pipeline, we consider each feature and split value when creating splits at each decision node. We consider tuning the hyperparameters in **Table B.1** within their corresponding search ranges. We consider the same search grid for both methods, as well as the same budget of hyperparameters. All experiments were performed on a virtual machine with 256 CPUs.

Model Selection. When training models for treatment effect estimation, we cannot observe the ground-truth performance on some held-out validation set to facilitate model selection. Thus, past work has considered *approximate* model selection techniques [182, 176, 74]. Such techniques choose hyperparameters by calculating a proxy metric on the validation dataset that may correlate with CATE estimation performance. However, the approximate nature of such techniques means that reported differences between approaches may be due more so to model selection than to the estimation approach. Throughout our experiments, we assume access to the ground-truth CATEs for choosing hyperparameters based on the maximum AUTOC in a held-out set. This setup controls for potential differences due to hyperparameter selection and allows for accurate comparisons of the proposed and baseline methods. As ground-truth performance estimates are not available in real applications, it remains imperative to improve the model selection challenge faced by all CATE estimation methods going forward. Random seeds and settings used will be available in the source code to enhance reproducibility.

B.5 Additional Results

Local AUTOC Maximization Splits: We compare our proposed method and baseline approach to building a decision tree that at any decision node M, maximizes the AUTOC in the sample S_M, rather than the full sample S. Note that this approach is not theoretically grounded towards the ultimate goal of maximizing AUTOC across the whole sample S, as

	N = 100	N = 250	N = 500	N = 1000
Proposed Performance	0.183 (0.100, 0.221)	**0.253 (0.197, 0.289)**	**0.292 (0.260, 0.316)**	0.326 (0.304, 0.338)
Local Split Performance	**0.195 (0.116, 0.221)**	0.236 (0.178, 0.280)	0.291 (0.240, 0.329)	**0.329 (0.301, 0.344)**
Baseline Performance	0.154 (0.075, 0.199)	0.223 (0.192, 0.242)	0.266 (0.221, 0.318)	0.323 (0.298, 0.347)

Table B.2: AUTOC performance on **Dataset 1**, comparing the proposed global splitting procedure, the local splitting procedure, and the baseline model. Splitting by maximizing AUTOC consistently outperforms the baseline model focused on accurate CATE estimation. Splitting based on local examples and global examples, however, results in similar performance.

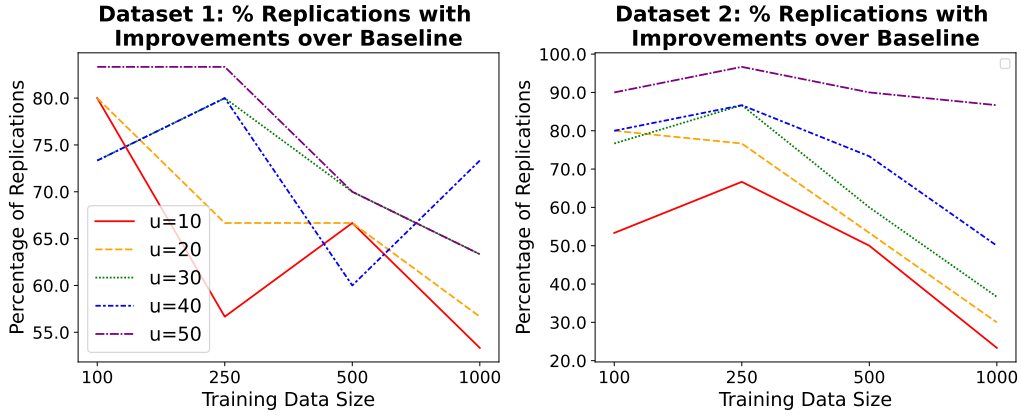

Figure B.1: The percentage of replications in which the proposed method outperforms the baseline in terms of ATE^u across different treatment thresholds u and training data size. The proposed method outperforms the baseline in up to $80-90\%$ of replications at different thresholds at low training data size, but the efficacy is only shown at higher treatment thresholds when enough training data is incorporated into the model.

a larger AUTOC for the subset S_M does not guarantee a larger AUTOC for the full sample S. However, this procedure provides a slightly faster proxy that may be considered for training. Results on **Dataset 1** can be found in **Table B.2**. Both techniques which split towards maximizing AUTOC result in better performance than the baseline method focused on accurate CATE estimation. However, the local and global splits tend to perform similarly. We hypothesize that this is due to the simplicity of our synthetic dataset. Empirically, local splits diverge from global splits at deeper levels of the decision trees, resulting in different estimators that achieve similar performance. As maximizing AUTOC in a local decision node does not guarantee the maximization of AUTOC across a whole sample, our proposed global splitting technique still provides a guarantee of maximizing our end goal. However, local splits may be used as a proxy for quicker training, despite the lack of theoretical guarantees.

Honest Decision Trees: We next show that our approach is amenable to the honest framework. We adapt both methods to the honest setting by using half of the training examples to create splits, and the other half to impute values. Decision trees with empty leaves for inference are ignored when aggregating results across the forest. We first report results on **Dataset 1** when using N = 250 training samples to train each method and evaluating on a held-out test set. The proposed method still outperforms the baseline method, achieving a median AUTOC of 0.259 (IQR: 0.206, 0.289) compared to 0.228 (IQR: 0.171, 0.256), and outperforming the baseline model on 28/30 replications. On **Dataset 2**, the proposed method continues to outperform the baseline technique at N = 250 training examples, achieving a median AUTOC of 0.735 (IQR: 0.511, 0.776) compared to a median AUTOC of 0.587 (0.397, 0.711) for the baseline method. The proposed method outperforms the baseline on a majority (29/30) of replications as well. Overall, these results show the ability of our method to be adapted to the honest setting, which may be preferred in settings where over-fitting is of great concern.

Comparison to Zhou et al. For completeness, we compare our proposed method with the loss function proposed by Zhou et al. implemented using a neural network [223]. We consider a small-sample regime with n = 250 training samples. To optimize the Zhou et al. loss function, we sweep over relevant hyperparameters such as the learning rate, the size of the neural network, and regularization strength. We find that in both synthetic datasets, our proposed method significantly outperforms this baseline technique as measured by the median AUTOC [IQR] on the test set (dataset 1: 0.088 [0.053-0.107] vs. 0.255 [0.185-0.279], dataset 2: 0.293 [0.216-0.378] vs. 0.750 [0.505-0.879]. Reweighting the loss function from past work using ground-truth propensity scores resulted in no improvement. We hypothesize that the poor performance is for two reasons. First, the loss function does not immediately transfer to the observational data setting due to confounding between the

treatment assignment and the outcomes. Second, our method directly optimizes for the value of the treatment policy at every threshold as measured by the AUTOC. Meanwhile, the method proposed by Zhou et al. relies on obtaining an unbiased estimate of the CATE to accurately rank scores. When CATEs cannot be estimated accurately, such as in low-data settings, methods to obtain unbiased CATEs may not lead to better AUTOC, as shown in Proposition 2.

Results at Specific Treatment Thresholds: To complement the ATE^u results in the main section, we first report the percentage of replications in which the proposed method outperforms the baseline in terms of ATE^u for different thresholds u (**Figure B.1**. At low training data sizes, the proposed method outperforms the baseline in over $80 - 90\%$ of replications across many thresholds, showing the efficacy of the proposed method. However, as more training data is incorporated, the baseline has the potential to slightly outperform the proposed method at low treatment thresholds, but the proposed method still performs well across a majority of settings. Next, report the TOC^u values for all $u \in [0, 1]$. These results can be found in **Figure B.2** for **Dataset 1**, and **Figure B.3** for **Dataset 2**. For both datasets, the efficacy of our proposed approach is better highlighted at lower data regimes. Across a majority of thresholds, our model consistently improves over random more than the baseline model does, as measured by TOC^u. At higher training data regimes, the efficacy of our model is more shown at treatment thresholds between $u = 30$ and $u = 50$. Moreover, our proposed method remains competitive with the baseline technique at higher data regimes, with only small drops in performance.

A Realistic Interpretation for Larger Data Regimes: We report the percentage of potential lives saved in our realistic set-up for higher training data regimes (N = 500, N = 1000) in **Figure B.4**. With $N = 500$ training data, the proposed method is still able to consistently improve upon the baseline technique. However, with $N = 1000$ examples used for training, our model begins to perform similarly, with only slight gains or losses compared to the baseline. This helps support our hypothesis that optimizing for AUTOC may improve upon the baseline in low training data regimes.

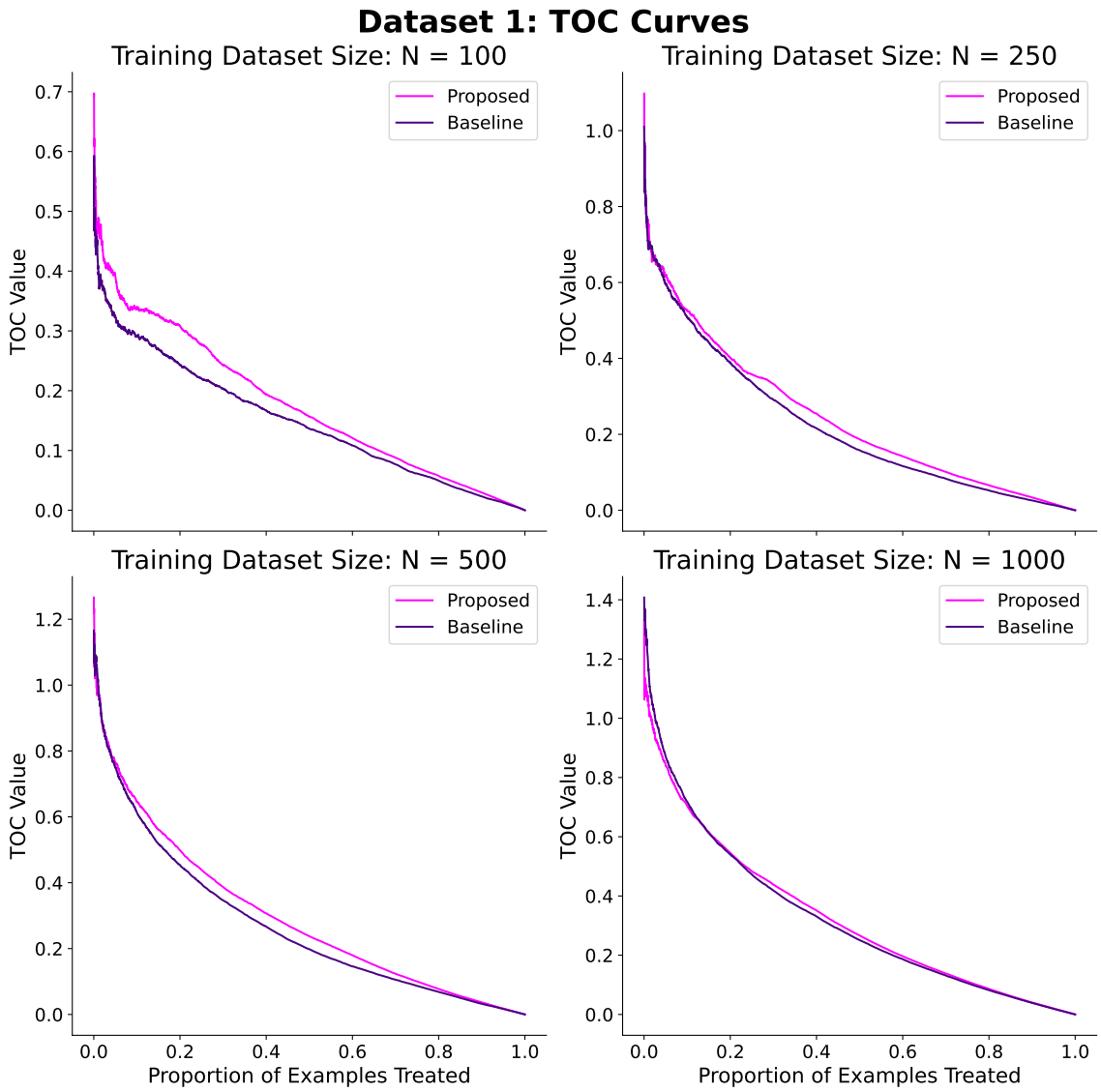

Figure B.2: TOC Curves for **Dataset 1**. In low-data settings, our method consistently results in a larger improvement in the ATE of the top percentage of individuals. As more data is included in our model, the improvements of our model are reduced, but our model still results in a larger TOC value across a majority of replications. When all individuals are treated, our method and the proposed method result in no improvement over random.

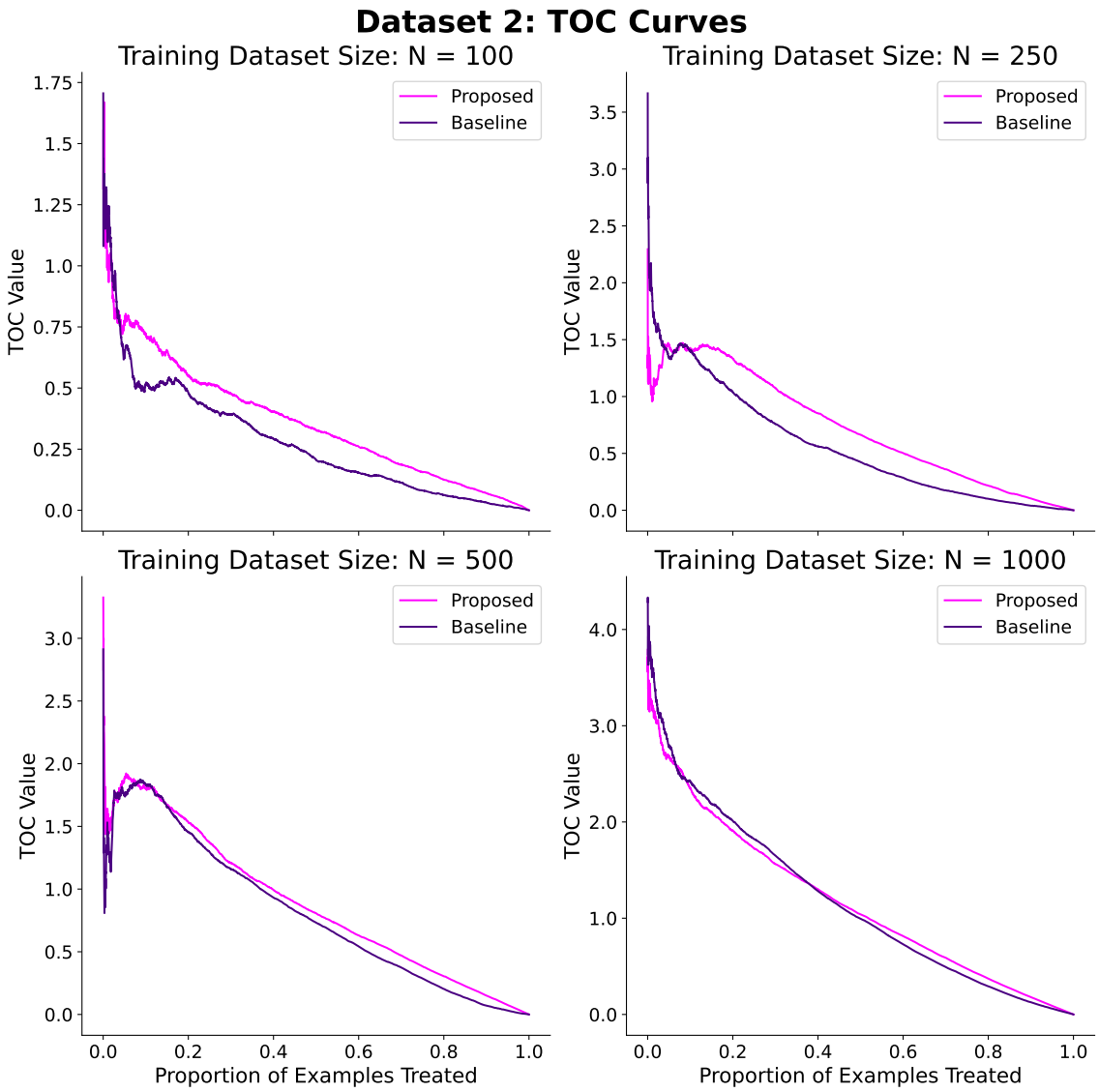

Figure B.3: TOC Curves for **Dataset 2**. In low data settings, our method results in a larger improvement in the ATE of the top percentage of individuals, particularly when the treatment threshold is above 10%. When $N = 1000$ data points are used to train the model, the baseline begins to slightly outperform the proposed method, especially at earlier treatment thresholds.

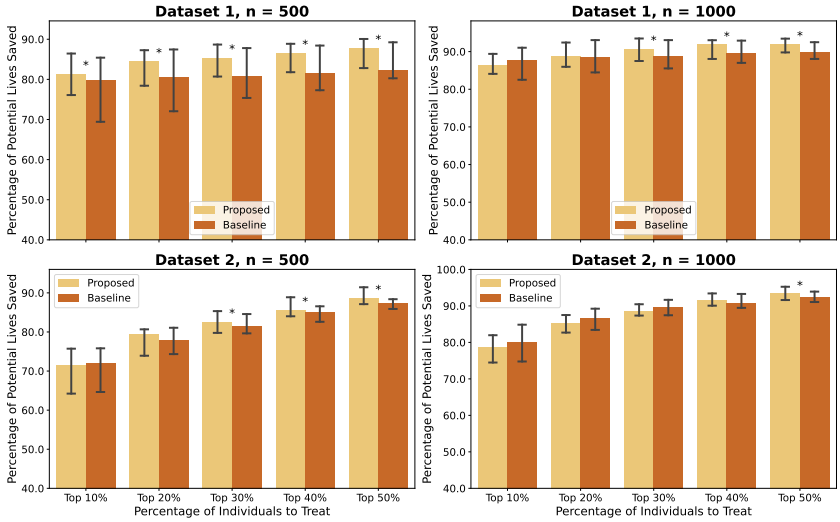

Figure B.4: Percentage of potential lives saved compared to the oracle across different treatment settings for high data settings for **Dataset 1** (top) and **Dataset 2** (bottom). Comparisons with asterisks represent scenarios in which the proposed method significantly outperforms the baseline technique as measured using a Wilcoxon signed rank test with a significance level of 0.05. At N = 500, the proposed method continues to perform well. However, as we add more training data, the models begin to perform similarly, with our model only performing slightly worse in some scenarios.

APPENDIX C

Appendix for Challenging Implicit Assumptions of Theory Through Empirical Evidence in CATE Estimation

C.1 Training Details

Training Setup. We implemented all neural network approaches in PyTorch. During training, we used the Adam optimizer [107]. Each model was given access to the same resources and was trained in a similar pipeline. We conducted this work on a server running Ubuntu 20.04.4 with 128 CPUs and 256 GB of RAM. We trained each model on a GeForce GTX 1080 Ti GPU for a large number of epochs. We performed early stopping during training for each model based on the training loss, with a patience level of 200 epochs. We performed a random search over all hyperparameters with a budget of 40 (**Table C.1**). We tuned the learning rate, weight decay, number of hidden layers, batch size, α, and other variables that weigh different components of a loss function (i.e., for matching and DragonNet), and whether to normalize the outputs during training. For all models, we kept the base architecture the same [182]. In line with past work a hidden size of 200 for the representation building layers, and a hidden size of 100 for the outcome layers. Moreover, we used *elu* as the non-linear activation function in the hidden layers. We follow past work and use all training examples to build nuisance estimates [43, 145]. In **Section C.3**, we confirm that this procedure outperforms a more traditional cross-fitting approach. For every hyperparameter configuration of a model, we use either a traditional feed-forward network (for conditional outcomes) or TARNet (for potential outcomes) with the same set of hyperparameters to learn these estimates. For all methods that require the use of the propensity score, unless otherwise noted, we estimated the propensity score using a regularized logistic regression model [15, 182, 128]. When using the propensity score, we clipped extremely low or high values to reduce variance and to ensure all values are between 0 and 1 when using noisy

propensity scores [12].

Model Selection. In traditional supervised learning, model selection or hyperparameter tuning (critical to training neural networks) often relies on measuring performance on a validation set. However, when training neural networks for treatment effect estimation, we cannot observe the ground-truth performance on some held-out validation set to facilitate model selection. Similar to **Chapter 4**, we assume access to ground-truth performance metrics for choosing hyperparameters to allow for an accurate comparison of different techniques away from issues due to model selection.

Table C.1: Hyperparameters and their corresponding search ranges.

Hyperparameter	**Hyperparameter Search Range**
Learning Rate	10^{-3}, 10^{-4}
Weight Decay	10^{-2}, 10^{-3}, 10^{-4}
Output Normalization	True, False
Number of Hidden Layers	4,5
Batch Size	50, 100, 200
α	.1, .5, 1.0, 2.0, 5.0, 10.0, 15.0, 20.0, 100.0

C.2 X-Learner Modification

We compare the X-Learner as proposed by [112], which uses two models to learn the direct CATE functions on the treatment and control group, to our proposed modification which uses a single multi-task model. On the synthetic dataset, the two models perform similarly, where the two-model version achieves a PEHE of 0.714 (SD: 0.222), and the single-model version achieves a PEHE of 0.711 (SD: 0.220). However, on the ACIC dataset, the single model version substantially outperforms the two model version, outperforming the baseline in 72 out of 77 DGPs (**Figure C.1**). By using a multi-task framework, all individuals can help learn a shared representation that is used by both CATE estimators. Throughout the main section, we thus considered the single-model version of the X-Learner.

C.3 Comparing Sample-Splitting to Using All Data

For models that require nuisance estimates (such as the R-Learner, DR-Learner, and X-Learner), we followed recent work and used all data for estimating these nuisance estimates [43, 145]. Cross-fitting and sample-splitting are necessary to ensure theoretical guarantees.

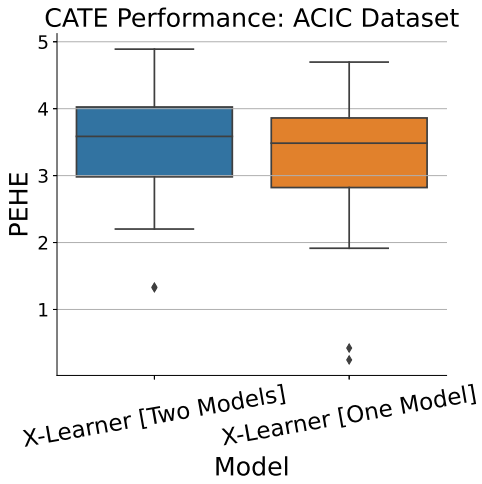

Figure C.1: CATE performance of different X-Learner models. The X-Learner using the TARNet architecture outperforms the traditional X-Learner proposed in [112] in 72 out of the 77 replications.

However, past work has found better empirical performance when using all data in finite and limited sample settings [43, 145]. To test this decision, we compare this approach to using cross-fitting to estimate the nuisance parameters, with K = 2 folds.

We compare the R-Learner, DR-Learner, and X-Learner to counterparts that use cross-fitting on the ACIC dataset. We consider the setting with ground-truth propensity scores and the potential outcomes need to be estimated. In **Figure C.2**, we find that models using all data outperform those that use cross-fitting to estimate the nuisance parameters, in line with recent work. Moreover, the general trends of the X-Learner outperforming all methods hold. Hence, due to the superior empirical performance, we continue with using all data to estimate the nuisance models.

C.4 Results for Omitted Techniques

Along with the methods considered in the main section, we implement and evaluate three more popular techniques.

First, we consider techniques that use the propensity score for adjustment during training. *Weighting Plug-In* is a popular direct learner that inversely weights the observed outcome using the propensity score [43, 99]. The CATE proxy learned in stage 1 is defined as $\hat{\tau}_i = y_i(\frac{t_i}{\hat{e}_i} - \frac{1-t_i}{1-\hat{e}_i})$, with $E[\hat{\tau}_i|\mathbf{x}_i] = \tau_i$. In stage 2, a neural network model is optimized to accurately estimate this proxy on the training set. *DragonNet* can also be trained such that it only uses the propensity score implicitly for adjustment during training [185]. By remov-

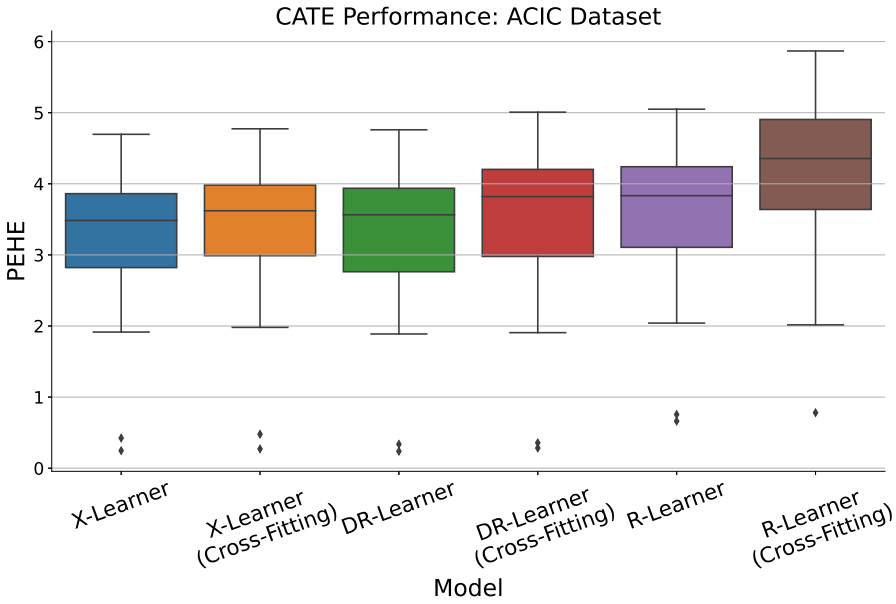

Figure C.2: CATE performance of different techniques with and without cross-fitting to estimate nuisance parameters. Techniques without cross-fitting outperform those that use cross-fitting to estimate the potential outcomes.

ing the targeted regularization term, the network is encouraged to learn a representation that is predictive of the treatment assignment and use this representation to learn potential outcomes. Throughout, we denote this method as *DragonNet*, and the method in the main section that includes the targeted regularization term as *DragonNet + tr*.

Finally, we consider a direct method that uses both propensity scores and outcome estimates during training. The *U-Learner* uses the same CATE proxy as the R-Learner built in stage 1 defined as $\tilde{\tau}_i = \frac{y_i - \hat{m}_i}{t_i - \hat{e}_i}$, where $E[\tilde{\tau}_i|\mathbf{x}_i] = \tau_i$ [142, 112]. Unlike the R-Learner, in stage 2, we regress the proxy outcome on the covariates without any weighting scheme. As studied in past work, the theoretical strengths of the R-Learner lie in the combination of the proxy outcome and the weighting scheme [55]. However, for completeness, we consider the performance of the U-Learner across all datasets.

We evaluate the performance of all techniques with access to ground-truth propensity scores. In the synthetic dataset, The weighting plug-in and the U-Learner perform poorly, despite access to ground-truth propensity scores (**Table C.2**). We hypothesize that this is likely due to the high variance induced by the inverse weighting of the propensity score in the proxy outcome, providing a poor estimate of the CATE to train with. Moreover, DragonNet without targeted regularization does not improve upon the full DragonNet, with both methods performing similarly.

Similar trends hold on the ACIC dataset (**Table C.3**). The U-Learner and the Weighting

Table C.2: Synthetic dataset results when using ground-truth propensity scores across all methods. The weighting plug-in and the U-Learner perform poorly, with the DR-Learner and the X-Learner still outperform all methods. Results in bold are statistically significant compared to TARNet.

Model	PEHE (SD) ↓	Improvements in PEHE ↑
TARNet	1.113 (0.201)	—
X-Learner	**0.711 (0.220)**	29
Weighting	**1.015 (0.161)**	20
Matching	**0.894 (0.163)**	24
Weighting Plug-In	1.606 (0.333)	3
U-Learner	2.224 (0.628)	0
R-Learner	**0.881 (0.139)**	26
DR-Learner	**0.714 (0.199)**	29
DragonNet + tr	**1.011 (0.255)**	20
DragonNet	1.137 (0.275)	15

Plug-In achieve the worst average ranks across all methods. However, DragonNet without targeted regularization performs better than with targeted regularization. Moreover, DragonNet without targeted regularization is able to provide improvements over TARNet. We hypothesize that this is because the combination of all loss functions can make training unstable, whereas training with only the outcome loss and the propensity score loss can provide more stable results while partially addressing confounding. However, the X-Learner still outperforms all techniques, such that the conclusions in the main section hold.

Table C.3: Top performing models and average rankings on ACIC 2016 across all methods with ground-truth propensity scores. DragonNet improves upon TARNet and DragonNet + tr, while the weighting plug-in and the U-Learner perform poorly.

Model	# Top-Performing PEHE ↑	Average Ranking PEHE ↓
TARNet	0/77	5.42/10
X-Learner	26/77	**2.23/10**
Weighting	1/77	4.17/10
Matching	**35/77**	3.30/10
Weighting Plug-In	0/77	8.74/10
U-Learner	0/77	9.92/10
R-Learner	0/77	7.09/10
DR-Learner	9/77	3.52/10
DragonNet + tr	0/77	6.09/10
DragonNet	6/77	4.52/10

APPENDIX D

Appendix for Mismatch in Sepsis Risk Stratification and Clinical Needs

D.1 Appendix for Mismatch Between Estimating Risk of Sepsis and Improving Patient Outcomes

D.1.1 Additional Cohort Details

We consider ICU admissions as evaluation cohorts in the main paper. In both evaluation sets, individuals may visit the ICU multiple times during the same stay, but in-hospital mortality is only defined once for an individual. For individuals with sepsis, we use the ICU stay in which they developed sepsis. For individuals without sepsis, we use their first ICU stay.

D.1.2 Features Extracted

For all patient admissions, we collect demographics, vital sign measurements, laboratory test results, and nursing score information, such as Glasgow coma scores and sedation information, throughout the hospitalization. In the U-M cohort, we also considered vital signs and comorbidities in encounters within the past year for making predictions. In the BIMDC cohort, we collect the same comorbidities for each admission using ICD-9 codes, filtering out conditions that are not present prior to the current hospitalization by using diagnosis-related groups [94].

We next describe the features extracted and used as input for all models. From demographic features, we collected age, race, ethnicity, marital status, and the source of admission. For vital signs, we included recorded heart rate, temperature, respiratory rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, and oxygen saturation. We

also included the following comorbidity information from past visits: cancer, chronic kidney disease, chronic liver disease, congestive heart failure, chronic obstructive pulmonary disease (COPD), diabetes, hypertension, obesity, dementia, drug abuse, and alcohol abuse. We included the following lab results as well: aspartate aminotransferase, bilirubin, creatinine, glucose, hematocrit, hemoglobin, international normalized ratio (INR), lymphocyte (absolute count and percentage), monocyte count, neutrophil level, platelet count, red blood cell count, segmented neutrophil count, urea nitrogen, white blood cell count, and albumin. Finally, we included nursing score information related to Glasgow coma score, pain scores, and sedation scores throughout a hospitalization. To the best of our abilities, we collected features in a similar way across the BIDMC and Michigan Medicine datasets. However, Information pertaining to red blood cell distribution width and advanced sedation information were unavailable in the BIDMC cohort. Moreover, comorbidities could not be captured in the same way across the datasets, as past visit information is not available in the BIDMC cohort.

D.1.3 Additional Model Training Details

When training XGBoost models, we tune the learning rate, the maximum depth of the trees, and the number of estimators used for the model. We select hyperparameters separately for each XGBoost model by maximizing performance using 5-fold cross-validation for each cohort consisting of a single window from a patient's admission, as described in the main section.

When training two-stage models such as the X-Learner and the DR-Learner, we use the S-Learner in the first stage to get estimates of mortality required for creating the proxy outcomes of the second stage. To obtain estimates of the propensity score, we follow our approach in the main section for building a model to estimate the risk of sepsis using an ensemble of XGBoost models. All models are built using the same training pipeline, selecting hyperparameters as described above.

D.1.4 Development Cohort

To train our machine learning models, we utilize held-out training sets that are distinct from the evaluation cohorts in the main section. We used 106,064 patient admissions for the U-M cohort and 13,864 ICU admissions for the BIDMC cohort. Of these admissions, 5,391 (5.1%) developed sepsis and 2,014 (1.9%) experienced in-hospital mortality in the U-M development cohort, while 1,108 (8.0%) developed sepsis during their stay and 1,231 (8.9%) experienced in-hospital mortality in the BIDMC development cohort.

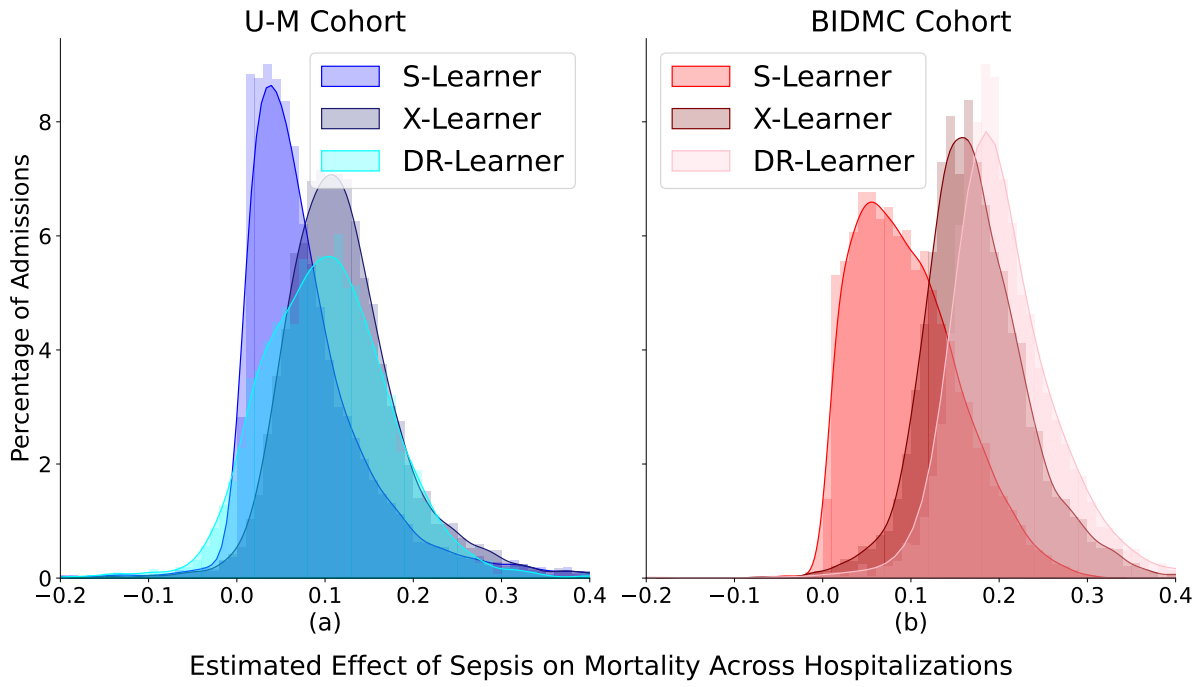

Figure D.1: Estimated effect of sepsis on mortality across hospital admissions and across different causal inference techniques. The estimated effect is once again on average positive in both datasets. Moreover, there is substantial heterogeneity in the estimated effect of sepsis on mortality regardless of the causal inference technique used to estimate these effects.

D.1.5 Statistical Analysis Results for All Models

Due to the strong performance of S-Learner, we reported results using this model in the main section. For completeness, we next report results using all methods.

The histogram of the estimated effect of sepsis on mortality confirms that the downstream effect is both positive and heterogeneous regardless of the causal inference technique employed (**Figure D.1**). In the U-M cohort, the S-Learner, X-Learner, and DR-Learner estimated a median effect of sepsis on mortality of 6.19 percentage points (entropy: 0.92), 11.34 percentage points (entropy: 1.08), and 10.00 percentage points (entropy: 1.17) respectively. In the BIDMC cohort, the three models estimated median effects of 8.82 percentage points (entropy: 0.87), 16.79 percentage points (entropy: 0.96), and 19.71 percentage points (entropy: 1.16).

The Spearman's correlation between the estimated risk of sepsis and the estimated effect of sepsis on mortality as measured by the S-Learner, X-Learner, and DR-Learner is 0.35 (95% CI: 0.33-0.37), 0.05 (95% CI: 0.02-0.07) and 0.30 (95% CI: 0.28-0.32) in the U-M cohort, and 0.31 (95% CI: 0.28-0.34), -0.24 (95% CI: -0.26- -0.21), and 0.04 (95% CI: 0.01-0.07) in the BIDMC cohort. Hence, almost all methods show a small positive relationship

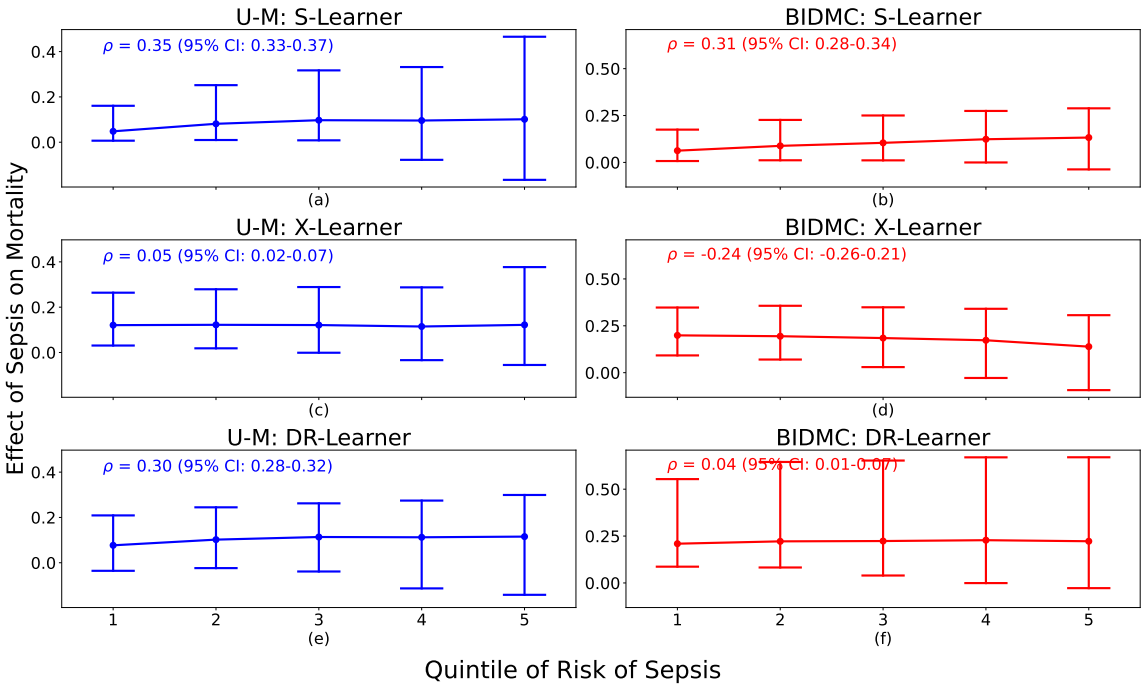

Figure D.2: The relationship between the effect of sepsis on mortality and the risk of developing sepsis across all causal inference techniques. Almost all methods estimate a slight positive relationship between the risk of sepsis and the severity of sepsis. There is large variance in the estimated effect of sepsis on mortality within windows with similar risks of sepsis.

between the risk of sepsis and the effect of sepsis on mortality, with the X-Learner applied to the BIDMC cohort being the only cohort that estimates a stronger negative relationship. Consistently, there is large variability in the effect of sepsis on mortality within patient windows with similar risks of sepsis across both datasets and across all causal inference techniques (**Figure D.2**). Meanwhile, as in the main analysis, there is a large group of windows that have a high estimated effect of sepsis on mortality but are at a low risk of sepsis and a large group of windows that are at a high risk of sepsis but with a low risk of mortality given the development of sepsis (**Figure D.3**).

D.1.6 Results for All Inpatients at U-M

For a fair comparison with BIDMC, we only focus our evaluation on ICU admissions at U-M in the main section. For completeness, we also report evaluation metrics on the full population of inpatients during the evaluation timeframe. For completeness, we also report evaluation metrics on the full population of inpatients during the timeframe of October 2018 to December 2020. In this analysis, we only remove patients from certain hospital

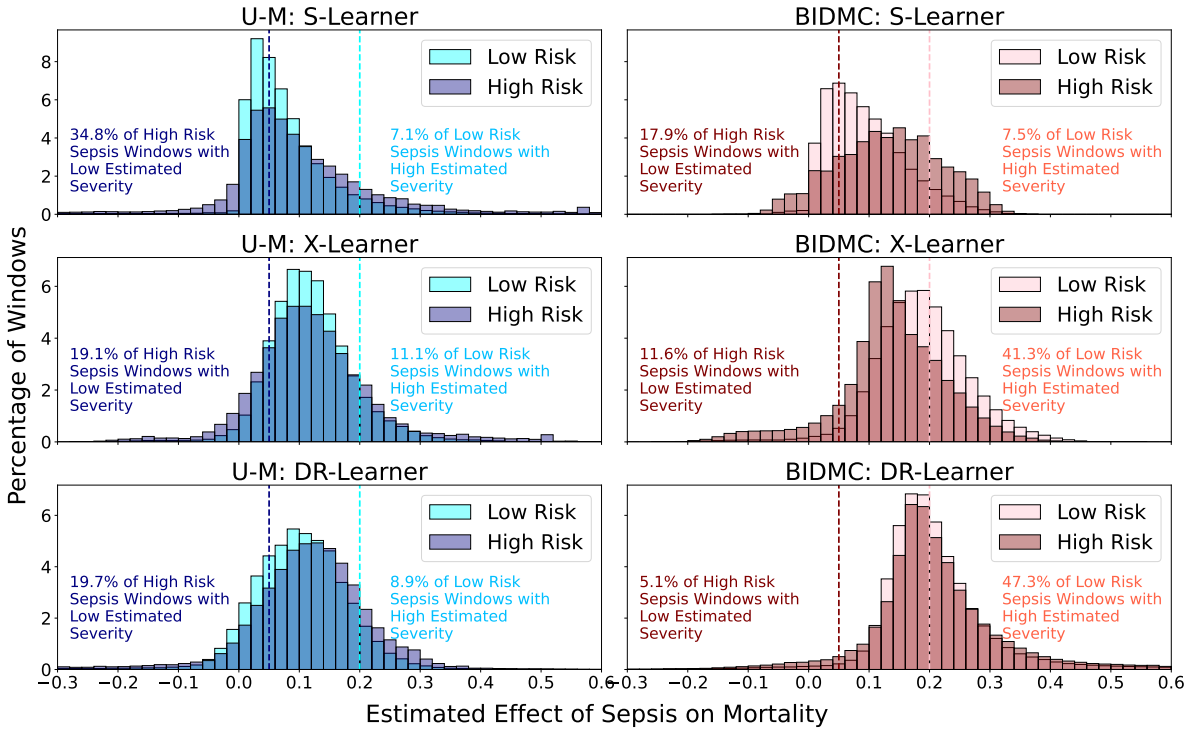

Figure D.3: The distribution of the severity of sepsis, as estimated by the effect of sepsis on mortality, is variable across both windows with high risk and low risk of sepsis. In all cohorts, as estimated by all models, there are windows with a high risk of sepsis whose development of sepsis would not adversely affect their likelihood of mortality. Meanwhile, there are also many low-risk windows whose risk of mortality would increase substantially if they were to develop sepsis.

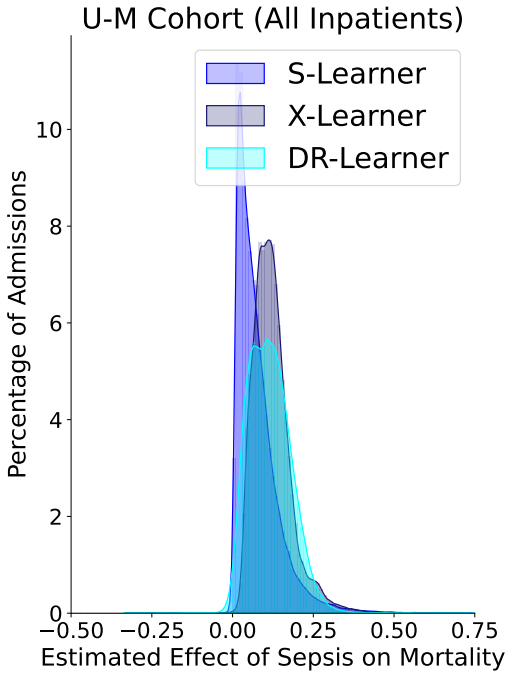

Figure D.4: Estimated effect of sepsis on mortality averaged over hospital admissions for the full set of inpatients at U-M. Similar to when focusing only on ICU patients, the estimated effect is on average positive and heterogeneous.

wards where sepsis is not a primary concern, such as patients admitted to the hospital for psychiatric or rehabilitation visits. This cohort consisted of 78,223 inpatient admissions, with 4,069 (5.2%) individuals developing sepsis, and 1,605 (2.1%) individuals experiencing in-hospital mortality. Of the septic individuals, 658 (16.2%) experienced in-hospital mortality. Meanwhile, 947 (1.3%) of the non-septic individuals experienced in-hospital mortality.

The machine learning model for estimating the risk of sepsis achieved an AUROC of 0.73 (95% CI: 0.73-0.74). The S-Learner achieved AUROCs of 0.94 (95% CI: 0.94-0.95) and 0.72 (95% CI: 0.71-0.75) for predicting in-hospital mortality for the no sepsis and sepsis populations. All causal inference techniques accurately estimated null treatment effects when performing the global null test, with the DR-Learner trained on septic individuals the only model achieving a non-zero mean-squared error (0.01 [95% CI: 0.01-0.01]).

The histogram of the averaged estimated effect of mortality on sepsis across patient admissions shows similar trends to when only focusing on ICU admissions as in the main paper (**Figure D.4**). The S-Learner, X-Learner, and DR-Learner estimated a median effect of sepsis on mortality of 5.93 percentage points (entropy: 0.77), 11.40 percentage points (entropy: 0.97), and 11.00 percentage points (entropy: 1.03) respectively in this cohort of admissions. The effect is positive and heterogeneous across all causal inference methods.

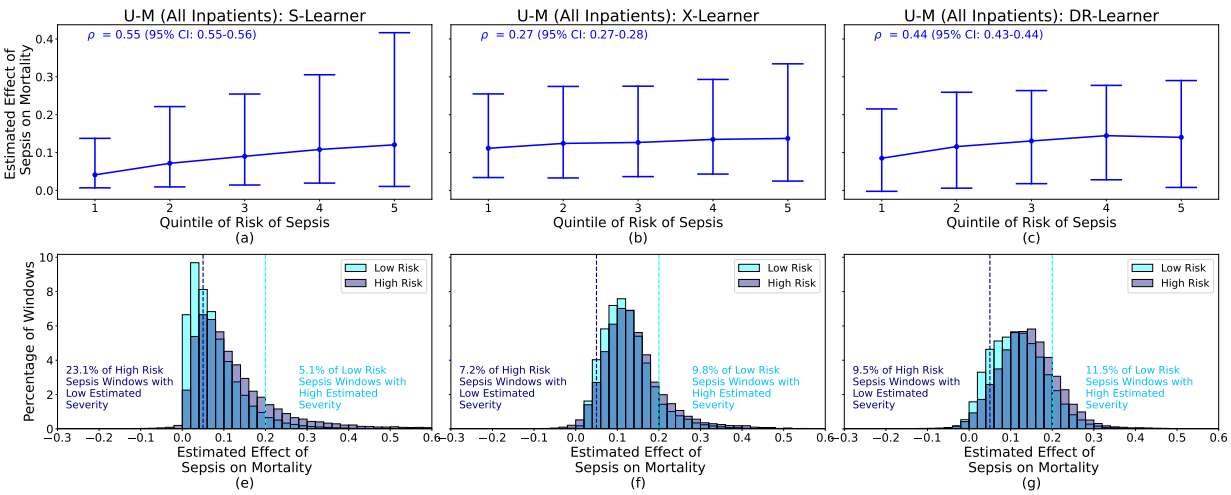

Figure D.5: Relationship between the effect of sepsis on mortality and the risk of developing sepsis (top) and the estimated effect of sepsis on mortality across different risk groups of developing sepsis (bottom) in all U-M inpatients. All methods show large variability in the effect of sepsis on mortality within individuals with similar risk of sepsis.

Finally, we visualize the relationship between the estimated effect of sepsis on mortality and the risk of sepsis (**Figure D.5**). Similar to the cohorts in the main paper, the trend is similar across all causal inference techniques, showing windows with a high risk of sepsis but a low estimated effect of sepsis on their mortality, and vice-versa. To view this further, we report the Spearman's correlation between the estimated risk of sepsis and the estimated effect of sepsis on mortality. As measured by the S-Learner, X-Learner, and DR-Learner, the correlations are 0.55 (95% CI: 0.55-0.56), 0.27 (95% CI: 0.27-0.28), and 0.44 (95% CI: 0.43-0.44). These moderate correlations show that the relationship between these two variables across all patient admissions and windows within an admission is not strong.

Bibliography

[1] What is sepsis? — sepsis — CDC. https://www.cdc.gov/sepsis/what-is-sepsis.html.

[2] Roy Adams, Katharine E Henry, Anirudh Sridharan, Hossein Soleimani, Andong Zhan, Nishi Rawat, Lauren Johnson, David N Hager, Sara E Cosgrove, Andrew Markowski, Eili Y Klein, Edward S Chen, Mustapha O Saheed, Maureen Henley, Sheila Miranda, Katrina Houston, Robert C Linton, Anushree R Ahluwalia, Albert W Wu, and Suchi Saria. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nature Medicine 2022 28:7*, 28(7):1455–1460, July 2022.

[3] Ahmed Alaa and Mihaela Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pages 129–138. PMLR, 2018.

[4] Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.

[5] Ahmed M Alaa and Mihaela van der Schaar. Deep multi-task gaussian processes for survival analysis with competing risks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2326–2334. Curran Associates Inc., 2017.

[6] Ahmad Alimadadi, Sachin Aryal, Ishan Manandhar, Patricia B Munroe, Bina Joe, and Xi Cheng. Artificial intelligence and machine learning to fight covid-19, 2020.

[7] Axel Andres, Aldo Montano-Loza, Russell Greiner, Max Uhlich, Ping Jin, Bret Hoehn, David Bigam, James Andrew Mark Shapiro, and Norman Mark Kneteman. A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis. *PloS one*, 13(3):e0193523, 2018.

[8] John E Angus. The probability integral transform and related results. *SIAM review*, 36(4):652–654, 1994.

[9] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.

[10] Eva Ascarza. Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, 55(1):80–98, 2018.

[11] Euan A Ashley. Towards precision medicine. *Nature Reviews Genetics*, 17(9):507–522, 2016.

[12] Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin Duke. Counterfactual representation learning with

balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pages 1972–1980. PMLR, 2021.

[13] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[14] Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

[15] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[16] Peter C Austin and Muhammad M Mamdani. A comparison of propensity score methods: a case-study estimating the effectiveness of post-ami statin use. *Statistics in medicine*, 25(12):2084–2106, 2006.

[17] Anand Avati, Tony Duan, Sharon Zhou, Kenneth Jung, Nigam H Shah, and Andrew Y Ng. Countdown regression: sharp and calibrated survival predictions. In *Uncertainty in Artificial Intelligence*, pages 145–155. PMLR, 2020.

[18] Laura S Bakosh, Renee M Snow, Jutta M Tobias, Janice L Houlihan, and Celestina Barbosa-Leiker. Maximizing mindful learning: Mindful awareness intervention improves elementary school students' quarterly grades. *Mindfulness*, 7(1):59–67, 2016.

[19] Brett K Beaulieu-Jones, William Yuan, Gabriel A Brat, Andrew L Beam, Griffin Weber, Marshall Ruffin, and Isaac S Kohane. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ digital medicine*, 4(1):62, 2021.

[20] Tellen D Bennett, Seth Russell, James King, Lisa Schilling, Chan Voong, Nancy Rogers, Bonnie Adrian, Nicholas Bruce, and Debashis Ghosh. Accuracy of the epic sepsis prediction model in a regional health system. February 2019.

[21] Artem Betlei, Eustache Diemert, and Massih-Reza Amini. Uplift modeling with generalization guarantees. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 55–65, 2021.

[22] Ioana Bica, Ahmed M Alaa, Craig Lambert, and Mihaela van der Schaar. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 109(1):87–100, 2021.

[23] Soorajnath Boominathan, Michael Oberst, Helen Zhou, Sanjat Kanjilal, and David Sontag. Treatment policy learning in multiobjective settings with fully observed outcomes. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1937–1947, 2020.

[24] Adam Booth, Angus Bruno Reed, Sonia Ponzo, Arrash Yassaee, Mert Aral, David Plans, Alain Labrique, and Diwakar Mohan. Population risk factors for severe disease and mortality in covid-19: A global systematic review and meta-analysis. *PLOS ONE*, 16:e0247461, 3 2021.

[25] Jean Bosco Sabuhoro, Bruno Larue, and Yvan Gervais. Factors determining the success or failure of canadian establishments on foreign markets: A survival analysis approach. *The International Trade Journal*, 20(1):33–73, 2006.

[26] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[27] Thomas C Brown. The concept of value in resource allocation. *Land economics*, 60(3):231–246, 1984.

[28] Luca Brunese, Francesco Mercaldo, Alfonso Reginelli, and Antonella Santone. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from x-rays. *Comput. Methods Programs Biomed.*, 196:105608, November 2020.

[29] Dan Burgin, Hollis R O'neal, Diana Hamer, Christopher B Thomas, and Tonya Jagneaux. ASSESSMENT OF EPIC SEPSIS PREDICTIVE ANALYTIC IMPACT ON ANTIBIOTIC USE IN THE ED. *Chest*, 162(4):A775, October 2022.

[30] Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

[31] Javier Enrique Camacho-Cogollo, Isis Bonet, Bladimir Gil, and Ernesto Iadanza. Machine learning models for early prediction of sepsis on large healthcare datasets. *Electronics 2022, Vol. 11, Page 1507*, 11:1507, 5 2022.

[32] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136, 2007.

[33] Alberto Caron, Gianluca Baio, and Ioanna Manolopoulou. Sparse bayesian causal forests for heterogeneous treatment effects estimation. *arXiv preprint arXiv:2102.06573*, 2021.

[34] Jared J Cash. Alert fatigue. *Am. J. Health. Syst. Pharm.*, 66(23):2098–2101, December 2009.

[35] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.

[36] David Cheng and Tianxi Cai. Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012*, 2021.

[37] 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.

[38] Zhixuan Chu, Stephen L Rathbun, and Sheng Li. Multi-task adversarial learning for treatment effect estimation in basket trials. In *Conference on Health, Inference, and Learning*, pages 79–91. PMLR, 2022.

[39] Richard Cookson, Christopher McCabe, and Aki Tsuchiya. Public healthcare resource allocation and the rule of rescue. *Journal of medical ethics*, 34(7):540–544, 2008.

[40] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

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

[42] Alicia Curth and Mihaela van der Schaar. Doing great at estimating cate? on the neglected assumptions in benchmark comparisons of treatment effect estimators. *arXiv preprint arXiv:2107.13346*, 2021.

[43] Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. *arXiv e-prints*, pages arXiv–2101, 2021.

[44] Ryan J. Delahanty, Jo Ann Alvarez, Lisa M. Flynn, Robert L. Sherwin, and Spencer S. Jones. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. *Annals of Emergency Medicine*, 73:334–344, 4 2019.

[45] Thomas Desautels, Jacob Calvert, Jana Hoffman, Qingqing Mao, Melissa Jay, Grant Fletcher, Chris Barton, Uli Chettipally, Yaniv Kerem, and Ritankar Das. Using transfer learning for improved mortality prediction in a data-scarce hospital setting. *Biomedical informatics insights*, 9:1178222617712994, 2017.

[46] Floris Devriendt, Jente Van Belle, Tias Guns, and Wouter Verbeke. Learning to rank for uplift modeling. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[47] Yealy Dm, Kellum Ja, Huang Dt, Barnato Ae, Weissfeld La, Pike F, Terndrup T, Wang He, Hou Pc, Lovecchio F, Filbin Mr, Shapiro Ni, and Angus Dc. A randomized trial of protocol-based care for early septic shock. *N. Engl. J. Med.*, 370(18):1683–1693, May 2014.

[48] Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, Dan Cervone, et al. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

[49] Tony Duan, Pranav Rajpurkar, Dillon Laird, Andrew Y Ng, and Sanjay Basu. Clinical value of predicting individual treatment effects for intensive blood pressure therapy: a machine learning experiment to estimate treatment effects from randomized trial data. *Circulation: Cardiovascular Quality and Outcomes*, 12(3):e005010, 2019.

[50] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[51] Carlos Fernández-Loría and Jorge Loría. Learning the ranking of causal effects with confounded data. *arXiv preprint arXiv:2206.12532*, 2022.

[52] Carlos Fernández-Loría and Foster Provost. Causal decision making and causal effect estimation are not the same... and why it matters. *INFORMS Journal on Data Science*, 2022.

[53] Ricard Ferrer, Antonio Artigas, David Suarez, Eduardo Palencia, Mitchell M Levy, Angel Arenzana, Xose Luis Pérez, and Josep Maria Sirvent. Effectiveness of treatments for severe sepsis. *https://doi.org/10.1164/rccm.200812-1912OC*, 180(9):861–866, December 2012.

[54] Simon R Finfer, Jean-Louis Vincent, Derek C Angus, and Tom Van Der Poll. Severe sepsis and septic shock. *https://doi.org/10.1056/NEJMra1208623*, 369(9):840–851, August 2013.

[55] Aaron Fisher. The connection between r-learning and inverse-variance weighting for estimation of heterogeneous treatment effects. *arXiv preprint arXiv:2307.09700*, 2023.

[56] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

[57] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

[58] FTM Freitas, AFOL Araujo, MIS Melo, and GAS Romero. Late-onset sepsis and mortality among neonates in a brazilian intensive care unit: a cohort study and survival analysis. *Epidemiology & Infection*, 147, 2019.

[59] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

[60] Fang Gao, Teresa Melody, Darren F Daniels, Simon Giles, and Samantha Fox. The impact of compliance with 6-hour and 24-hour sepsis bundles on hospital mortality in patients with severe sepsis: a prospective observational study. *Crit. Care*, 9(6):R764, 2005.

[61] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.

[62] Marzyeh Ghassemi, Marco Pimentel, Tristan Naumann, Thomas Brennan, David Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[63] Thomas A Glass, Steven N Goodman, Miguel A Hernán, and Jonathan M Samet. Causal inference in public health. *Annual review of public health*, 34:61–75, 2013.

[64] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[65] Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

[66] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[67] Mark Goldstein, Xintian Han, Aahlad Puli, Adler Perotte, and Rajesh Ranganath. X-cal: Explicit calibration for survival analysis. *Advances in Neural Information Processing Systems*, 33, 2020.

[68] Lalla Aïda Guindo, Monika Wagner, Rob Baltussen, Donna Rindress, Janine van Til, Paul Kind, and Mireille M Goetghebeur. From efficacy to equity: Literature review of decision criteria for resource allocation and healthcare decisionmaking. *Cost effectiveness and resource allocation*, 10(1):1–13, 2012.

[69] Fethi Gül, Mustafa Kemal Arslantaş, İsmail Cinel, and Anand Kumar. Changing definitions of sepsis. *Turkish Journal of Anaesthesiology and Reanimation*, 45(3):129, 2017.

[70] Pierre Gutierrez and Jean-Yves Gérardy. Causal inference and uplift modelling: A review of the literature. In *International conference on predictive applications and APIs*, pages 1–13. PMLR, 2017.

[71] P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

[72] Humza Haider. *MTLR: Survival Prediction with Multi-Task Logistic Regression*, 2019. R package version 0.2.1.

[73] Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85):1–63, 2020.

[74] Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5880–5887, 2019.

[75] Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020.

[76] Tobias Hatt, Jeroen Berrevoets, Alicia Curth, Stefan Feuerriegel, and Mihaela van der Schaar. Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv preprint arXiv:2202.12891*, 2022.

[77] Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (TREWScore) for septic shock. *Sci. Transl. Med.*, 7(299), August 2015.

[78] Miguel A Hernan and James M Robins. Causal inference, 2020.

[79] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[80] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[81] Chang Hu, Lu Li, Weipeng Huang, Tong Wu, Qiancheng Xu, Juan Liu, and Bo Hu. Interpretable machine learning for early prediction of prognosis in sepsis: a discovery and validation study. *Infectious Diseases and Therapy*, 11(3):1117–1132, 2022.

[82] Muhammad Ibrahim and Mark Carman. Comparing pointwise and listwise objective functions for random-forest-based learning-to-rank. *ACM Transactions on Information Systems (TOIS)*, 34(4):1–38, 2016.

[83] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

[84] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[85] Kosuke Inoue, Susan Athey, and Yusuke Tsugawa. Machine-learning-based high-benefit approach versus conventional high-risk approach in blood pressure management. *International Journal of Epidemiology*, page dyad037, 2023.

[86] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[87] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[88] Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal effect inference failure with uncertainty-aware models. *arXiv e-prints*, pages arXiv–2007, 2020.

[89] Andrew Jesson, Panagiotis Tigas, Joost van Amersfoort, Andreas Kirsch, Uri Shalit, and Yarin Gal. Causal-bald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data. *Advances in Neural Information Processing Systems*, 34:30465–30478, 2021.

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

[91] Fredrik D Johansson, Nathan Kallus, Uri Shalit, and David Sontag. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.

[92] Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

[93] Alistair E.W. Johnson, Jerome Aboab, Jesse D. Raffa, Tom J. Pollard, Rodrigo O. Deliberato, Leo A. Celi, and David J. Stone. A comparative analysis of sepsis identification methods in an electronic database. *Critical care medicine*, 46:494, 2018.

[94] Alistair E.W. Johnson, David J. Stone, Leo A. Celi, and Tom J. Pollard. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25:32–39, 1 2018.

[95] Annahieta Kalantari, Haney Mallemat, and Scott D Weingart. Sepsis definitions: The search for gold and what CMS got wrong. *West. J. Emerg. Med.*, 18(5):951, August 2017.

[96] Nathan Kallus. Classifying treatment responders under causal effect monotonicity. In *International Conference on Machine Learning*, pages 3201–3210. PMLR, 2019.

[97] Nathan Kallus, Xiaojie Mao, and Madeleine Udell. Causal inference with noisy and missing covariates via matrix factorization. In *Advances in neural information processing systems*, pages 6921–6932, 2018.

[98] Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.

[99] Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in neural information processing systems*, pages 9269–9279, 2018.

[100] Fahad Kamran, Shengpu Tang, Erkin Otles, Dustin S McEvoy, Sameh N Saleh, Jen Gong, Benjamin Y Li, Sayon Dutta, Xinran Liu, Richard J Medford, Thomas S Valley, Lauren R West, Karandeep Singh, Seth Blumberg, John P Donnelly, Erica S Shenoy, John Z Ayanian, Brahmajee K Nallamothu, Michael W Sjoding, and Jenna Wiens. Early identification of patients admitted to hospital for covid-19 at risk of clinical deterioration: model development and multisite external validation study. *BMJ*, 376(11), February 2022.

[101] Fahad Kamran and Jenna Wiens. Estimating calibrated individualized survival curves with deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 240–248, 2021.

[102] Sanjat Kanjilal, Michael Oberst, Sooraj Boominathan, Helen Zhou, David C Hooper, and David Sontag. A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. *Science Translational Medicine*, 12(568):eaay5067, 2020.

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

[104] 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):24, 2018.

[105] Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

[106] Norawit Kijpaisalratana, Daecha Sanglertsinlapachai, Siwapol Techaratsami, Khrongwong Musikatavorn, and Jutamas Saoraya. Machine learning algorithms for early sepsis detection in the emergency department: A retrospective study. *International Journal of Medical Informatics*, 160, 4 2022.

[107] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. iclr (2015), 2015.

[108] Eike-Henner W Kluge. Resource allocation in healthcare: implications of models of medicine as a profession. *Medscape General Medicine*, 9(1):57, 2007.

[109] Michael C Knaus, Michael Lechner, and Anthony Strittmatter. Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, 24(1):134–161, 2021.

[110] William A Knaus, Frank E Harrell, Joanne Lynn, Lee Goldman, Russell S Phillips, Alfred F Connors, Neal V Dawson, William J Fulkerson, Robert M Califf, Norman Desbiens, et al. The support prognostic model: objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203, 1995.

[111] Pekka Korhonen and Mikko Syrjänen. Resource allocation based on efficiency analysis. *Management Science*, 50(8):1134–1144, 2004.

[112] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10):4156, 2019.

[113] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

[114] Simon Meyer Lauritsen, Bo Thiesson, Marianne Johansson Jørgensen, Anders Hammerich Riis, Ulrick Skipper Espelund, Jesper Bo Weile, and Jeppe Lange. The framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards. *npj Digital Medicine 2021 4:1*, 4(1):1–12, November 2021.

[115] Changhee Lee, William Zame, Ahmed Alaa, and Mihaela Schaar. Temporal quilting for survival analysis. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 596–605, 2019.

[116] 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.

[117] Mitchell M Levy, Laura E Evans, and Andrew Rhodes. The surviving sepsis campaign bundle: 2018 update. *Intensive Care Med.*, 44(6):925–928, June 2018.

[118] Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

[119] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[120] Yaobin Ling, Pulakesh Upadhyaya, Luyao Chen, Xiaoqian Jiang, and Yejin Kim. Heterogeneous treatment effect estimation using machine learning for healthcare application: tutorial and benchmark. *arXiv preprint arXiv:2109.12769*, 2021.

[121] Vincent Liu, Gabriel J Escobar, John D Greene, Jay Soule, Alan Whippy, Derek C Angus, and Theodore J Iwashyna. Hospital deaths in patients with sepsis from 2 independent cohorts. *JAMA*, 312(1):90–92, July 2014.

[122] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[123] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

[124] Margaux Luck, Tristan Sylvain, Héloïse Cardinal, Andrea Lodi, and Yoshua Bengio. Deep learning for patient-specific kidney graft survival analysis. *arXiv preprint arXiv:1705.10245*, 2017.

[125] P G Lyons, B Ramsey, M Simkins, and T M Maddox. TP014 DIAGNOSTIC AND SCREENING INSIGHTS IN PULMONARY, CRITICAL CARE, AND SLEEP / thematic poster session how useful is the epic sepsis prediction model for predicting sepsis?

[126] Patrick G Lyons, Mackenzie R Hofford, Sean C Yu, Andrew P Michelson, Philip R O Payne, Catherine L Hough, and Karandeep Singh. Factors associated with variability in the performance of a proprietary sepsis prediction model across 9 networked hospitals in the US. *JAMA Intern. Med.*, April 2023.

[127] Anil N Makam, Oanh K Nguyen, and Andrew D Auerbach. Diagnostic accuracy and effectiveness of automated electronic sepsis alert systems: A systematic review. *J. Hosp. Med.*, 10(6):396–402, June 2015.

[128] Maggie Makar, Fredrik Johansson, John Guttag, and David Sontag. Estimation of bounds on potential outcomes for decision making. In *International Conference on Machine Learning*, pages 6661–6671. PMLR, 2020.

[129] Ben J Marafino, Alejandro Schuler, Vincent X Liu, Gabriel J Escobar, and Mike Baiocchi. Predicting preventable hospital readmissions with causal machine learning. *Health services research*, 55:993–1002, 2020.

[130] Florian B Mayr, Sachin Yende, and Derek C Angus. Epidemiology of severe sepsis. *https://doi.org/10.4161/viru.27372*, 5(1):4–11, 2013.

[131] Aditya Krishna Menon, Brendan Van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent corruption. *arXiv preprint arXiv:1605.00751*, 2016.

[132] Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

[133] Rupert G Miller Jr. *Survival analysis*. John Wiley & Sons, 2011.

[134] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[135] Xenia Miscouridou, Adler Perotte, Noémie Elhadad, and Rajesh Ranganath. Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, pages 244–256, 2018.

[136] Michael Moor, Max Horn, Bastian Rieck, Damian Roqueiro, and Karsten Borgwardt. Early recognition of sepsis with gaussian process temporal convolutional networks and dynamic time warping. *Proceedings of Machine Learning Research*, 106:1, 2019.

[137] Allan H Murphy. A new vector partition of the probability score. *Journal of applied Meteorology*, 12(4):595–600, 1973.

[138] Andrea Naghi and Christian P Wirths. Finite sample evaluation of causal machine learning methods: Guidelines for the applied researcher. 2021.

[139] Preetam Nandy, Xiufan Yu, Wanjun Liu, Ye Tu, Kinjal Basu, and Shaunak Chatterjee. Generalized causal tree for uplift modeling. *arXiv preprint arXiv:2202.02416*, 2022.

[140] Engineering National Academies of Sciences, Medicine, et al. *Framework for equitable allocation of COVID-19 vaccine*. National Academies Press, 2020.

[141] Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*, 2020.

[142] Xinkun Nie and Stefan Wager. Learning objectives for treatment effect estimation. *arXiv preprint arXiv:1712.04912*, 2017.

[143] Kenneth R Niswander and Myron Gordon. *The women and their pregnancies: the Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke*. National Institute of Health, 1972.

[144] J Oh, M Makar, C Fusco, and Others. A generalizable, Data-Driven approach to predict daily risk of clostridium difficile infection at two large academic health centers first authors of equal contribution. b senior authors of equal contribution. *Infect. Control Hosp. Epidemiol.*, 39(4):425–433, 2018.

[145] Gabriel Okasa. Meta-learners for estimation of causal effects: Finite sample cross-fit performance. *arXiv preprint arXiv:2201.12692*, 2022.

[146] Diego Olaya, Jonathan Vásquez, Sebastián Maldonado, Jaime Miranda, and Wouter Verbeke. Uplift modeling for preventing student dropout in higher education. *Decision Support Systems*, 134:113320, 2020.

[147] Brent C Opmeer. Electronic health records as sources of research data. *Jama*, 315(2):201–202, 2016.

[148] Harsh Parikh, Carlos Varjao, Louise Xu, and Eric Tchetgen Tchetgen. Validating causal inference methods. In *International Conference on Machine Learning*, pages 17346–17358. PMLR, 2022.

[149] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala.

Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

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

[151] Steven D Pinkerton, Ana P Johnson-Masotti, Arthur Derse, and Peter M Layde. Ethical issues in cost-effectiveness analysis. *Evaluation and program planning*, 25(1):71–83, 2002.

[152] Robi Polikar. Ensemble learning. *Ensemble machine learning: Methods and applications*, pages 1–34, 2012.

[153] Nicholas Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, pages 14–21, 2007.

[154] Pranav Rajpurkar, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017.

[155] Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. *arXiv preprint arXiv:1608.02158*, 2016.

[156] Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. Deep recurrent survival analysis. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[157] Chanu Rhee, Raymund Barretto Dantes, Lauren Epstein, and Michael Klompas. Using objective clinical data to track progress on preventing and treating sepsis: CDC's new 'adult sepsis event' surveillance strategy. *BMJ Qual. Saf.*, 28(4):305–309, April 2019.

[158] Chanu Rhee, Travis M Jones, Yasir Hamad, Anupam Pande, Jack Varon, Cara O'Brien, Deverick J Anderson, David K Warren, Raymund B Dantes, Lauren Epstein, and Michael Klompas. Prevalence, underlying causes, and preventability of Sepsis-Associated mortality in US acute care hospitals. *JAMA Network Open*, 2(2):e187571–e187571, February 2019.

[159] Chanu Rhee, Zilu Zhang, Sameer S Kadri, David J Murphy, Greg S Martin, Elizabeth Overton, Christopher W Seymour, Derek C Angus, Raymund Dantes, Lauren Epstein, David Fram, Richard Schaaf, Rui Wang, and Michael Klompas. Sepsis surveillance using adult sepsis events simplified eSOFA criteria versus sepsis-3 SOFA criteria. *Crit. Care Med.*, 47(3):307, March 2019.

[160] Andrew Rhodes, Laura E Evans, Waleed Alhazzani, Mitchell M Levy, Massimo Antonelli, Ricard Ferrer, Anand Kumar, Jonathan E Sevransky, Charles L Sprung, Mark E Nunnally, Bram Rochwerg, Gordon D Rubenfeld, Derek C Angus, Djillali Annane, Richard J Beale, Geoffrey J Bellinghan, Gordon R Bernard, Jean Daniel

Chiche, Craig Coopersmith, Daniel P De Backer, Craig J French, Seitaro Fujishima, Herwig Gerlach, Jorge Luis Hidalgo, Steven M Hollenberg, Alan E Jones, Dilip R Karnad, Ruth M Kleinpell, Younsuk Koh, Thiago Costa Lisboa, Flavia R Machado, John J Marini, John C Marshall, John E Mazuski, Lauralyn A McIntyre, Anthony S McLean, Sangeeta Mehta, Rui P Moreno, John Myburgh, Paolo Navalesi, Osamu Nishida, Tiffany M Osborn, Anders Perner, Colleen M Plunkett, Marco Ranieri, Christa A Schorr, Maureen A Seckel, Christopher W Seymour, Lisa Shieh, Khalid A Shukri, Steven Q Simpson, Mervyn Singer, B Taylor Thompson, Sean R Townsend, Thomas Van der Poll, Jean Louis Vincent, W Joost Wiersinga, Janice L Zimmerman, and R Phillip Dellinger. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock: 2016. *Intensive Care Medicine 2017 43:3*, 43(3):304–377, January 2017.

[161] Severi Rissanen and Pekka Marttinen. A critical look at the consistency of causal estimation with deep latent variable models. *Advances in Neural Information Processing Systems*, 34:4207–4217, 2021.

[162] Emanuel Rivers, Bryant Nguyen, Suzanne Havstad, Julie Ressler, Alexandria Muzzin, Bernhard Knoblich, Edward Peterson, and Michael Tomlanovich. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N. Engl. J. Med.*, 345(19):1368–1377, November 2001.

[163] Iain Robertson-Steel. Evolution of triage systems. *Emergency medicine journal*, 23(2):154–155, 2006.

[164] Joshua A Rolnick and Gary E Weissman. Early warning systems: The neglected importance of timing. *J. Hosp. Med.*, 14(7):445–447, July 2019.

[165] 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.

[166] Paul R Rosenbaum and Donald B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

[167] Joachim Roski, George W Bo-Linn, and Timothy A Andrews. Creating value in health care through big data: opportunities and policy implications. *Health affairs*, 33(7):1115–1122, 2014.

[168] Casey Ross. [no title]. https://www.statnews.com/2021/09/27/epic-sepsis-algorithm-antibiotics-model/. Accessed: 2023-5-30.

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

[170] Cynthia Rudin and Robert E Schapire. Margin-based ranking and an equivalence between adaboost and rankboost. 2009.

[171] Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32:303–327, 2012.

[172] Suchi Saria and Katharine E Henry. Too many definitions of sepsis: Can machine learning leverage the electronic health record to increase accuracy and bring consensus? *Crit. Care Med.*, pages 137–141, February 2020.

[173] Aaron L Sarvet, Kerollos N Wanis, Jessica Young, Roberto Hernandez-Alejandro, Miguel A Hernán, and Mats J Stensrud. Causal inference with limited resources: proportionally-representative interventions. *arXiv preprint arXiv:2002.11846*, 2020.

[174] Michiel Schinkel, Tom van der Poll, and W Joost Wiersinga. Artificial intelligence for early sepsis detection – a word of caution. *Am. J. Respir. Crit. Care Med.*, April 2023.

[175] Luregn J Schlapbach, Niranjan Kissoon, Abdulelah Alhawsawi, Maha H Aljuaid, Ron Daniels, Luis A Gorordo-Delsol, Flavia Machado, Imrana Malik, Emmanuel Fru Nsutebu, Simon Finfer, and Konrad Reinhart. World sepsis day: a global agenda to target a leading cause of morbidity and mortality. *https://doi.org/10.1152/ajplung.00369.2020*, 319(3):L518–L522, September 2020.

[176] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018.

[177] David LB Schwappach. Resource allocation, social values and the qaly: a review of the debate and empirical evidence. *Health Expectations*, 5(3):210–222, 2002.

[178] Mark P Sendak, William Ratliff, Dina Sarro, Elizabeth Alderton, Joseph Futoma, Michael Gao, Marshall Nichols, Mike Revoir, Faraz Yashar, Corinne Miller, Kelly Kester, Sahil Sandhu, Kristin Corey, Nathan Brajer, Christelle Tan, Anthony Lin, Tres Brown, Susan Engelbosch, Kevin Anstrom, Madeleine Clare Elish, Katherine Heller, Rebecca Donohoe, Jason Theiling, Eric Poon, Suresh Balu, Armando Bedoya, and Cara O'Brien. Real-World integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Med Inform*, 8(7):e15182, July 2020.

[179] William R Shadish and Peter M Steiner. A primer on propensity score analysis. *Newborn and Infant Nursing Reviews*, 10(1):19–26, 2010.

[180] Nilay D Shah, Ewout W Steyerberg, and David M Kent. Big data and predictive analytics: recalibrating expectations. *Jama*, 320(1):27–28, 2018.

[181] K Shailaja, B Seetharamulu, and MA Jabbar. Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*, pages 910–914. IEEE, 2018.

[182] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

[183] Supreeth P Shashikumar, Christopher S Josef, Ashish Sharma, and Shamim Nemati. DeepAISE—an interpretable and recurrent neural survival model for early prediction of sepsis. *Artif. Intell. Med.*, 113:102036, March 2021.

[184] Supreeth P Shashikumar, Gabriel Wardi, Atul Malhotra, and Shamim Nemati. Artificial intelligence sepsis prediction algorithm learns to say "i don't know". *NPJ Digit Med*, 4(1):134, September 2021.

[185] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pages 2503–2513, 2019.

[186] Yue Shi, Martha Larson, and Alan Hanjalic. List-wise learning to rank with matrix factorization for collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 269–272, 2010.

[187] Pannagadatta K Shivaswamy, Wei Chu, and Martin Jansche. A support vector approach to censored targets. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 655–660. IEEE, 2007.

[188] Mervyn Singer, Clifford S. Deutschman, Christopherwarren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tomvan Der Poll, Jean Louis Vincent, and Derek C. Angus. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA - Journal of the American Medical Association*, 315:801–810, 2 2016.

[189] Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

[190] Harald Steck, Balaji Krishnapuram, Cary Dehing-Oberije, Philippe Lambin, and Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. *Advances in neural information processing systems*, 20, 2007.

[191] Ewout W Steyerberg et al. *Clinical prediction models*. Springer, 2019.

[192] Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *J. Am. Med. Inform. Assoc.*, 27(12):1921–1934, December 2020.

[193] Yasir Tarabichi, Aurelia Cheng, David Bar-Shain, Brian M McCrate, Lewis H Reese, Charles Emerman, Jonathan Siff, Christine Wang, David C Kaelber, Brook Watts, and Michelle T Hecker. Improving timeliness of antibiotic administration using a provider and pharmacist facing sepsis early warning system in the emergency department setting: A randomized controlled quality improvement initiative. *Crit. Care Med.*, 50(3):418–427, March 2022.

[194] Ambuj Tewari and Sougata Chaudhuri. Generalization error bounds for learning to rank: Does the length of document lists matter? In *International Conference on Machine Learning*, pages 315–323. PMLR, 2015.

[195] Adrienne Torda. Ethical issues in pandemic planning. *Medical journal of Australia*, 185(S10):S73–S76, 2006.

[196] Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, and Ewout W Steyerberg. Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7, 2019.

[197] Ben Van Calster and Andrew J Vickers. Calibration of risk prediction models: impact on decision-analytic performance. *Medical decision making*, 35(2):162–169, 2015.

[198] Arjun K Venkatesh, Todd Slesinger, Jessica Whittle, Tiffany Osborn, Emily Aaronson, Craig Rothenberg, Nalani Tarrant, Pawan Goyal, Donald M Yealy, and Jeremiah D Schuur. Preliminary performance on the new CMS sepsis-1 national quality measure: Early insights from the emergency quality network (E-QUAL). *Ann. Emerg. Med.*, 71(1):10–15.e1, January 2018.

[199] Cameron Voloshin, Hoang M Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

[200] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[201] Lu Wang, Yan Li, Jiayu Zhou, Dongxiao Zhu, and Jieping Ye. Multi-task survival analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 485–494. IEEE, 2017.

[202] Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):110, 2019.

[203] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, and Tie-Yan Liu. A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, volume 8, page 6, 2013.

[204] R Scott Watson and Joseph A Carcillo. Scope and epidemiology of pediatric sepsis. *Pediatr. Crit. Care Med.*, 6(3 SUPPL.), May 2005.

[205] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

[206] Jenna Wiens and Erica S Shenoy. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases*, 66(1):149–153, 2018.

[207] Julian M Williams, Jaimi H Greenslade, Juliet V McKenzie, Kevin Chu, Anthony F T Brown, and Jeffrey Lipman. Systemic inflammatory response syndrome, quick sequential organ function assessment, and organ dysfunction: Insights from a prospective database of ED patients with infection. *Chest*, 151(3):586–596, March 2017.

[208] Andrew Wong, Erkin Otles, John P Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.*, 181(8):1065–1070, August 2021.

[209] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199, 2008.

[210] Yizhe Xu, Katelyn Bechler, Alison Callahan, and Nigam Shah. Principled estimation and evaluation of treatment effect heterogeneity: A case study application to dabigatran for patients with atrial fibrillation. *Journal of Biomedical Informatics*, page 104420, 2023.

[211] Yizhe Xu, Nikolaos Ignatiadis, Erik Sverdrup, Scott Fleming, Stefan Wager, and Nigam Shah. Treatment heterogeneity with survival outcomes. In *Handbook of Matching and Weighting Adjustments for Causal Inference*, pages 445–482. Chapman and Hall/CRC, 2023.

[212] Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966*, 2021.

[213] Meicheng Yang, Xingyao Wang, Hongxiang Gao, Yuwen Li, Xing Liu, Jianqing Li, and Chengyu Liu. Early prediction of sepsis using multi-feature fusion based xgboost learning and bayesian optimization. In *The IEEE Conference on Computing in Cardiology (CinC)*, volume 46, pages 1–4, 2019.

[214] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *arXiv preprint arXiv:2002.02770*, 2020.

[215] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

[216] Jeffrey Yuk-Chiu Yip. Healthcare resource allocation in the covid-19 pandemic: Ethical considerations from the perspective of distributive justice within public health. *Public Health in Practice*, 2:100111, 2021.

[217] Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in Neural Information Processing Systems*, pages 1845–1853, 2011.

[218] Morteza Zabihi, Serkan Kiranyaz, and Moncef Gabbouj. Sepsis prediction in intensive care unit using ensemble of xgboost models. In *2019 Computing in Cardiology (CinC)*, pages Page–1. IEEE, 2019.

[219] Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10923–10930, 2021.

[220] Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Learning overlapping representations for the estimation of individualized treatment effects. *arXiv preprint arXiv:2001.04754*, 2020.

[221] Xin Zhao, Wenqian Shen, Guanjun Wang, et al. Early prediction of sepsis based on machine learning algorithm. *Computational Intelligence and Neuroscience*, 2021, 2021.

[222] Yan Zhao, Xiao Fang, and David Simchi-Levi. Uplift modeling with multiple treatments and general response types. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 588–596. SIAM, 2017.

[223] Hao Zhou, Shaoming Li, Guibin Jiang, Jiaqi Zheng, and Dong Wang. Direct heterogeneous causal learning for resource allocation problems in marketing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5446–5454, 2023.

[224] José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.