Learning to Rank for Optimal Treatment Allocation Under Resource Constraints

Fahad Kamran
University of Michigan

Maggie Makar
University of Michigan

Jenna Wiens
University of Michigan

Abstract

Current causal inference approaches for estimating conditional average treatment effects (CATEs) often prioritize accuracy. However, in resource constrained settings, decision makers may only need a ranking of individuals based on their estimated CATE. In these scenarios, exact CATE estimation may be an unnecessarily challenging task, particularly when the underlying function is difficult to learn. In this work, we study the relationship between CATE estimation and optimizing for CATE ranking, demonstrating that optimizing for ranking may be more appropriate than optimizing for accuracy in certain settings. Guided by our analysis, we propose an approach to directly optimize for rankings of individuals to inform treatment assignment that aims to maximize benefit. Our tree-based approach maximizes the expected benefit of the treatment assignment using a novel splitting criteria. In an empirical case-study across synthetic datasets, our approach leads to better treatment assignments compared to CATE estimation methods as measured by expected total benefit. By providing a practical and efficient approach to learning a CATE ranking, this work offers an important step towards bridging the gap between CATE estimation techniques and their downstream applications.

1 INTRODUCTION

The problem of resource allocation or prioritizing interventions is common across various fields (Brown, 1984; Korhonen and Syrjänen, 2004; National Academies of Sciences et al., 2020; Cookson et al., 2008). In healthcare, for instance, clinicians must triage patients for different levels of care (Robertson-Steel, 2006). In marketing, companies must prioritize customers for marketing campaigns and retention programs (Ascarza, 2018; Radcliffe, 2007). Similarly, in education, targeted interventions can lower dropout rates or improve academic performance (Bakosh et al., 2016; Olaya et al., 2020). While numerous other examples exist, in this work, we use 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 (Kluge, 2007; Guindo et al., 2012). For example, clinicians may prioritize monitoring and additional care for a subset of individuals at risk of deteriorating due to sepsis (Filbin et al., 2018). This problem setting is especially relevant during a global pandemic (Jöbges et al., 2020), but even prior to the pandemic healthcare systems around the world were already strained with long wait times and burnt out clinicians (Dzau et al., 2018). Accordingly, in some settings, clinicians may be forced to triage patients. These triaging decisions may be based, at least in part, on a ranking of who is likely to benefit most from a particular intervention (i.e., the treatment effect) (Kluge, 2007; Schwappach, 2002; Yadlowsky et al., 2021; Inoue et al., 2023). Tools that could help clinicians in estimating benefit from observational data could help in assisting clinicians in defining this ranking. However, estimating treatment effects from observational data is rarely straightforward.

Conditional average treatment effects (CATEs) quantify the effect of a treatment on an outcome given an individual's covariates. However, estimating CATEs using observational data is challenging due to potential confounding (Foster et al., 2011; Hernan and Robins, 2020). Accordingly, past research has worked to improve accuracy and sample efficiency in CATE estima-

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

tion through novel machine learning techniques (Glass et al., 2013; Alaa and van der Schaar, 2017; Shalit et al., 2017; Wager and Athey, 2018; Hernan and Robins, 2020; Hassanpour and Greiner, 2020; Zhang et al., 2020; Kennedy, 2020). 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, framing a new problem of causal classification for identifying treatment responders (Kallus, 2019; Athey and Wager, 2021; Fernández-Loría and Provost, 2022). In these settings, the goal is to learn whether an individual will benefit from treatment, as defined by some threshold on the estimated CATE, and prioritize treatment for these individuals. Past work has both studied the disconnect between this problem and CATE estimation and has studied methods for directly optimizing for this usecase. In this work, we build upon this recent paradigm shift and extend this idea beyond a binary classification problem. Specifically, we study study the problem of optimal ranking policies without the need for an *a priori* threshold to treat, similar to triage. As thresholds for determining whom to treat may vary depending on the application, and may change even within the same application, we need approaches that are agnostic to a particular threshold and provide an overall ranking.

There are strong parallels between these causal inference tasks and the field of reinforcement learning. Our problem setting of interest could be framed as a bandit problem in which a model estimates a ranking to maximize overall benefit, rather than the standard approaches of measuring the value of different treatment policies (akin to causal effect estimation) or learning treatment assignments for each example (similar to causal classification). We study the potential for a model-based approach that directly optimizes for maximizing overall benefit in comparison to these standard approaches.

Recent research in the field of uplift modelling has begun to study this problem (Rzepakowski and Jaroszewicz, 2012; Betlei et al., 2021; Zhao et al., 2017; Zhou et al., 2023). For example, Zhou et al. (2023) propose a new objective function that does not focus on the accuracy of the CATE estimates to obtain unbiased CATE estimates that may be used to rank individuals for resource allocation. While related, past work assumes access to data from a randomized controlled trial or with binary outcomes. These differences in the problem setting change 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.

We study the disconnect between the problem of optimizing for optimal treatment allocation and unbiased CATE estimation, which is often an objective of past work (Zhou et al., 2023). Building on recent work, we focus on a theoretical and empirical exploration of the disconnect between these two problem setups. 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 (Cao et al., 2007).

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 optimal ranking, i.e., one that maximizes benefit across all treatment thresholds. In this paper, 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 a ranking that maximizes overall benefit. Based on our findings, we develop a novel tree-based approach that produces a ranking of individuals that maximizes expected benefit across all treatment thresholds. We show that our approach, in low sample settings, is more sample-efficient and outperforms CATE estimation techniques that focus on accuracy. 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.
- Empirically, we explore the potential for directly maximizing expected benefit compared to optimizing for CATE accuracy. Across a range of settings with limited data, our approach is more sample-efficient and outperforms methods that focus primarily on accurate CATE estimation in low-data regimes.

2 PROBLEM SET-UP AND BACKGROUND

Setup. We study a setting where the decision maker aims to identify the top $u\%$ of individuals who will benefit most from some intervention, for all u. 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 (Rubin, 1974; Splawa-Neyman et al., 1990). Specifically, for an individual 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 $Y_i(1)$ (Pearl, 2009). We define the CATE as: $\tau_i = CATE(\mathbf{x}_i) = \mathbb{E}[y|\mathbf{x}_i, do(t = 1)] - \mathbb{E}[y|\mathbf{x}_i, do(t = 0)] = Y_i(1) - Y_i(0)$.

Goal. To identify the top $u\%$ of individuals who will benefit (i.e., have the greatest CATE) for all u, 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 (Rudin and Schapire, 2009; Steck et al., 2007). In our setting, in addition to the pairwise truth function, we also have ground-truth continuous 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. Given a ranking, we aim to measure the overall benefit from treatment across all possible thresholds. To measure the expected benefit of treating the top $u\%$ of patients in sample S, as identified by model f, we assume that the CATE τ_i is observed and may be used for evaluation. Formally, we define $D^u_S(f)$ as the top $u\%$ of individuals ranked by the model, i.e., $D^u_S(f) = \{i | f(\mathbf{x}_i) \geq \psi(\{f(\mathbf{x}_i)_{i \in S}\}, u)\}$, where $\psi(a, u)$ is the u^{th} percentile of the empirical distribution of a. The average benefit from treatment for these individuals is defined as $ATE^u_S(f) = \frac{1}{|D^u_S(f)|}\sum_{i \in D^u_S(f)} \tau_i$. A larger $ATE^u_S(f)$ 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., $TOC^u_S(f) = ATE^u_S(f) - \frac{1}{|S|}\sum_{k=1}^{|S|} \tau_k$ (Yadlowsky et al., 2021). A value of 0 represents no improvement over random. Finally, to measure this across all treatment thresholds u, we use the *Area Under the TOC* (AUTOC). For an arbitrary function f and a sample S,

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

(Yadlowsky et al., 2021). 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) (Yadlowsky et al., 2021). Hence, calculating the AUTOC aligns directly with the goal of measuring the overall benefit from treatment if a model is used to triage examples across every possible threshold. 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 be negative 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 (Yadlowsky et al., 2021).

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 (Shalit et al., 2017; Hernan and Robins, 2020; Imbens and Rubin, 2015). We discuss the implications of these assumptions in the conclusion.

3 THEORETICAL ANALYSIS

In this section, we study the relationship between accurate CATE estimation and optimal ranking for maximizing overall benefit defined based on the treatment effect (**Figure 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.

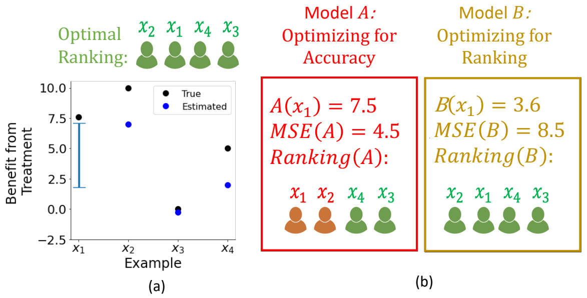

Figure 1: A motivating example. Consider four individuals, and a model that has estimated CATEs for individuals $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$. To achieve better mean-squared-error (MSE), the model should predict a value close to the true CATE (7.5). However, the model can achieve an optimal ranking by estimating the CATE of the remaining example ($\mathbf{x}_1$) anywhere in a large interval between the estimates of $\mathbf{x}_2$ and $\mathbf{x}_4$. 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 $f : \mathcal{X} \rightarrow \mathbb{R}$ 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 $f : \mathcal{X} \rightarrow \mathbb{R}$ 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 can be found in **Appendix B**. **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 an $L^M(f)$ greater than 0. More generally, any function $f \in \mathcal{F}$ that biases each example i by less than half its margin γ_i is guaranteed to 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 minimum 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 that optimizes AUTOC will be just as large, if not larger, than the set of solutions that optimizes CATE accuracy. Maximizing 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 $f : \mathcal{X} \rightarrow \mathbb{R}$ 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 is found in **Appendix B**. 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 op-

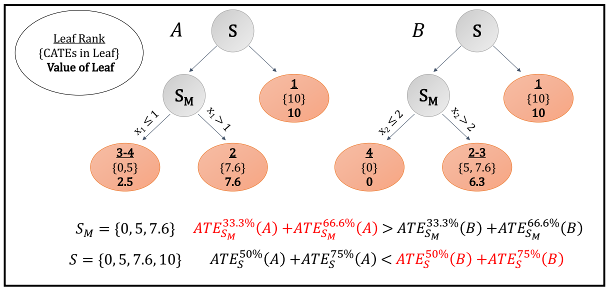
Leaf Rank
{CATEs in Leaf}
Value of Leaf
A
B
S
S_M
x_1 ≤ 1
x_1 > 1
x_2 ≤ 2
x_2 > 2
1 {10} 10
3-4 {0,5} 2.5
2 {7.6} 7.6
4 {0} 0
2-3 {5, 7.6} 6.3
$S_M = \{0, 5, 7.6\}$ $ATE_{S_M}^{33.3\%}(A) + ATE_{S_M}^{66.6\%}(A) > ATE_{S_M}^{33.3\%}(B) + ATE_{S_M}^{66.6\%}(B)$
$S = \{0, 5, 7.6, 10\}$ $ATE_S^{50\%}(A) + ATE_S^{75\%}(A) < ATE_S^{50\%}(B) + ATE_S^{75\%}(B)$

Figure 2: The importance of global splits. We define a subtree with data S, in which we aim to split at decision node M, resulting in either tree A or B. A 'local' split based on only data in S_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.

timizing 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 and especially when the margin γ is large and easy to learn, where estimating the CATE function exactly might be challenging, but optimizing AUTOC could be easier. We test this hypothesis empirically and seek approaches that optimize for AUTOC directly in **Section 4**.

Though we focus on the AUTOC, our major theoretical results apply to other ranking-related evaluation metrics, which can be viewed as modifications to the AUTOC (e.g., Qini curve and AUPEC) (Yadlowsky et al., 2021; Imai and Li, 2023). For all of these metrics, our main propositions hold, such that accurate CATEs are not required for accurate ranking, and better CATE performance does not guarantee better ranking. However, we continue with the AUTOC, and leave empirical explorations of other ranking metrics for future work.

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 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 within a sample S.

Optimizing For and Calculating AUTOC. Maximizing AUTOC for a sample S is 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 (Ibrahim and Carman, 2016).

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

1. Assign a score $T(\mathbf{x}_i)$ to each individual i in S based on the average outcome of the leaf node in which $\mathbf{x}_i$ falls.

2. Calculate $AUTOC_S(T)$ using the scores $T(\mathbf{x}_i)$. To handle ties where multiple examples have the same predicted score, average across all possible orderings to simulate breaking ties at random (Yadlowsky et al., 2021).

Learning Decision Trees to Maximize AUTOC. We propose an approach for building a tree T to optimize for AUTOC. To aid in our explanation, 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. Standard regression trees choose 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 potential limitations of local splits (**Figure 2**). While all data are considered at each split, the tree is still grown greedily, thus computation time increases only slightly (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 (Ibrahim and Carman, 2016). Given this training procedure, we bootstrap our data multiple times and build many decision trees to overcome overfitting and improve performance, as in standard random forest (Breiman, 2001). 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 (Chernozhukov et al., 2018; Kennedy, 2020). The nuisance parameters $\hat{m}$ and $\hat{e}$ represent nonparametric estimates of the ground-truth propensity score and potential outcome functions. Under our assumptions, $E[\tilde{\tau}_i|\mathbf{x}_i] \rightarrow \tau_i$ as $n \rightarrow \infty$. 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 (Yadlowsky et al., 2021). These doubly robust proxies can be built using cross-fitting. Then, when making a split at decision node M, we find the k, v pair that maximizes $\widetilde{AUTOC}_S(T^M_{k,v})$. A model that directly maximizes the AUTOC using the doubly robust proxy will also, in expectation, maximize the true AUTOC.

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 with large margins between examples. 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.

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 (Kennedy, 2020). The doubly robust proxy $\tilde{\tau}_i$ for each example can only be built for individuals for whom treatments and outcomes are observed. Hence, at inference time, on a new set of examples for whom the treatment and outcome is 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 standard 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. To give all methods the best opportunity to learn, we use cross-fitting with decision trees to estimate the potential outcomes and 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. We tune the same hyperparameters for both methods using the same search space. We tune 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 D** for more details and set-up) [1].

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 evaluate without

[1] Code can be found at `https://github.com/MLD3/Learning-to-Rank-for-Optimal-Treatment-Allocation-Under-Resource-Constraints`

ground truth. Hence, as a first step, in this work 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 (Athey and Wager, 2021; Caron et al., 2021). Specifically, we generate two datasets. In **Dataset 1**, the ground truth τ_i function is built to create different groups of individuals with different treatment effects, resulting in large margins on average between individuals.

Dataset 1

$$
\begin{aligned}
&\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_{10x10}),\\
&t_i|\mathbf{x}_i \sim Bern(\frac{1}{1+e^{-x_{i,3}}}),\\
&\epsilon_i|\mathbf{x}_i, t_i \sim \mathcal{N}(0,1),\\
&\tau_i|\mathbf{x}_i{=}((x_{i,1})_+ + (x_{i,2})_+ - 1)/2,\\
&y_i|\mathbf{x}_i, \tau_i, \epsilon_i, t_i{=}\max(0, x_{i,3}+x_{i,4}) + t_i\tau_i + \epsilon_i
\end{aligned}
$$

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.

Dataset 2

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

Though semi-synthetic causal inference datasets have been studied in the past, we use fully synthetic datasets to control every portion of the data generating process as a first step for validating the proposed method. This decision is supported by recent work calling into question the use of common benchmark datasets, such as the IHDP and ACIC 2016 dataset, for comparing treatment effect models (Curth and van der Schaar, 2021). For example, the IHDP dataset violates the overlap assumption necessary for causal effect estimation and inherently favors some techniques over others. Moreover, other semi-synthetic datasets, such as the TCGA dataset are not immediately applicable to our setting with binary treatments. Hence, we use synthetic data to provide a better understanding of the potential of the proposed methodology.

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 healthcare, many diseases are rare and many patient populations have less representation in the data. Due to this, many problems in the field of healthcare are plagued with issues due to a limited number of examples (Desautels et al., 2017; Chen et al., 2021). 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 E**.

5.2 Results

AUTOC Performance: 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 3**). As the sample size increases, both approaches 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 (i.e., only maximizing AUTOC using data in the current splitting node) (**Appendix E**). 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 (Athey and Imbens, 2016). 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

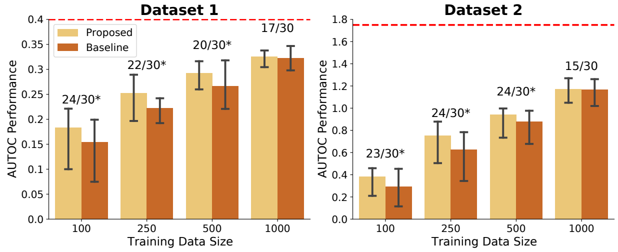

Figure 3: Median and IQR AUTOC, and how many times the proposed method outperforms the baseline across 30 replications. Asterisks show where the proposed method significantly outperforms the baseline technique as measured with the Wilcoxon signed rank test ($\alpha = 0.05$). The maximum AUTOC achievable is indicated by the red dashed line. At low sample sizes, the proposed method outperforms the baseline.

show that our model still outperforms the baseline technique in a low-sample setting (**Appendix E**). Finally, for completeness, we also compare our approach to that of Zhou et al. (2023) in **Appendix E** and show that our proposed approach significantly outperforms this baseline.

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**). Across training set sizes 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. In larger 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 consistently outperforms the baseline at thresholds beyond $u = 50$ (**Appendix E**).

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

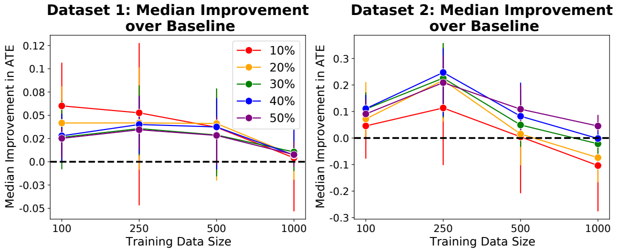

Figure 4: The median and IQR of improvement of the proposed approach over the baseline in ATE^u across thresholds u. Our model excels across treatment thresholds at low-data settings, despite not being trained for a particular treatment threshold. With more training data ($N = 1000$), our approach is more efficacious when treating a larger fraction of individuals.

In both datasets, the proposed method outperforms the baseline in terms of % lives saved (**Figure 5**). At $u = 30$, the proposed method 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%). This trend is consistent across all thresholds, with the proposed method constantly outperforming the baseline technique, with up to a 6.4% increase. In a setting where 2,000 lives could be saved by allocating treatments, an improvement of 6.4% means saving 128 additional lives over the baseline. In data-rich settings, the proposed method matches the performance of the baseline or performs slightly worse (**Appendix E**). Overall, this evaluation demonstrates the potential for direct maximization of expected benefit in resource-constrained settings.

6 DISCUSSION AND CONCLUSION

In this work, 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

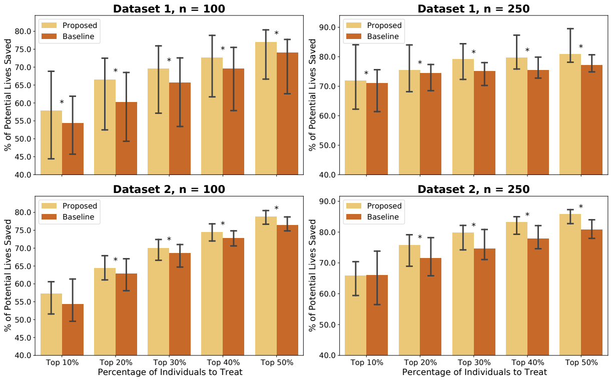

Figure 5: Median and IQR of the percentage of potential lives saved compared to the oracle across different thresholds in low data settings for **Dataset 1** (top) and **Dataset 2** (bottom). Asterisks represent scenarios in which the proposed method significantly outperforms the baseline technique as 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.

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 (Fernández-Loría and Provost, 2022; Athey and Wager, 2021; Kallus, 2019). 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 data. 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 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 (Torda, 2006; Pinkerton et al., 2002). For example, there exist many ethical constraints which may be considered when allocating interventions, as shown during the COVID-19 pandemic (Yip, 2021). We emphasize that a ranking based on benefit alone is not intended to automate clinical decisions, but merely inform decisions. Our work is intended to study one tool that may be used to augment this decision-making, which may also be combined with other societal considerations. Similarly, we emphasize that we do not advocate for ignoring accuracy and precision of treatment effects in *all* clinical settings. Accurate treatment effects are especially important in precision health (e.g., cancer treatment), where patient-specific treatment effect estimates are essential to guide decision-making (Kent et al., 2018). In this work, we study a situation in which patient-specific estimates are not necessary, and study alternatives to these approaches in such settings. In addition, like most 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 optimal ranking 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 optimal rankings (Fernández-Loría and Loría, 2022). 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 for 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 using observational data. Overall, our work represents an important step for bridging theory and practice of resource allocation techniques.

Acknowledgments

This work was supported by Cisco Research. and the National Science Foundation (NSF award nos. IIS 2124127 and 2153083). The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Cisco Research. or the National Science Foundation. We also thank the anonymous reviewers and members of the MLD3 lab for their valuable feedback.

References

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

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

Assaad, S., Zeng, S., Tao, C., Datta, S., Mehta, N., Henao, R., Li, F., and Carin, L. (2021). Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pages 1972–1980. PMLR.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Bakosh, L. S., Snow, R. M., Tobias, J. M., Houlihan, J. L., and Barbosa-Leiker, C. (2016). Maximizing mindful learning: Mindful awareness intervention improves elementary school students' quarterly grades. *Mindfulness*, 7(1):59–67.

Betlei, A., Diemert, E., and Amini, M.-R. (2021). Uplift modeling with generalization guarantees. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 55–65.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

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

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. (2007). Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136.

Caron, A., Baio, G., and Manolopoulou, I. (2021). Sparse bayesian causal forests for heterogeneous treatment effects estimation. *arXiv preprint arXiv:2102.06573*.

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., and Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Cookson, R., McCabe, C., and Tsuchiya, A. (2008). Public healthcare resource allocation and the rule of rescue. *Journal of medical ethics*, 34(7):540–544.

Curth, A. and van der Schaar, M. (2021). Doing great at estimating cate? on the neglected assumptions in benchmark comparisons of treatment effect estimators. *arXiv preprint arXiv:2107.13346*.

Desautels, T., Calvert, J., Hoffman, J., Mao, Q., Jay, M., Fletcher, G., Barton, C., Chettipally, U., Kerem, Y., and Das, R. (2017). Using transfer learning for improved mortality prediction in a data-scarce hospital setting. *Biomedical informatics insights*, 9:1178222617712994.

Devriendt, F., Van Belle, J., Guns, T., and Verbeke, W. (2020). Learning to rank for uplift modeling. *IEEE Transactions on Knowledge and Data Engineering*.

Dzau, V. J., Kirch, D. G., Nasca, T. J., et al. (2018). To care is human—collectively confronting the clinician-burnout crisis. *N Engl J Med*, 378(4):312–314.

Fernández-Loría, C. and Loría, J. (2022). Learning the ranking of causal effects with confounded data. *arXiv preprint arXiv:2206.12532*.

Fernández-Loría, C. and Provost, F. (2022). Causal decision making and causal effect estimation are not the same... and why it matters. *INFORMS Journal on Data Science*.

Filbin, M. R., Thorsen, J. E., Lynch, J., Gillingham, T. D., Pasakarnis, C. L., Capp, R., Shapiro, N. I., Mooncai, T., Hou, P. C., Heldt, T., et al. (2018). Challenges and opportunities for emergency department sepsis screening at triage. *Scientific reports*, 8(1):11059.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880.

Glass, T. A., Goodman, S. N., Hernán, M. A., and Samet, J. M. (2013). Causal inference in public health. *Annual review of public health*, 34:61–75.

Guindo, L. A., Wagner, M., Baltussen, R., Rindress, D., van Til, J., Kind, P., and Goetghebeur, M. M. (2012). From efficacy to equity: Literature review of decision criteria for resource allocation and healthcare decisionmaking. *Cost effectiveness and resource allocation*, 10(1):1–13.

Gutierrez, P. and Gérardy, J.-Y. (2017). Causal inference and uplift modelling: A review of the literature. In *International conference on predictive applications and APIs*, pages 1–13. PMLR.

Hassanpour, N. and Greiner, R. (2019). Counterfactual regression with importance sampling weights. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5880–5887.

Hassanpour, N. and Greiner, R. (2020). Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*.

Hernan, M. A. and Robins, J. M. (2020). Causal inference.

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

Imai, K. and Li, M. L. (2023). Experimental evaluation of individualized treatment rules. *Journal of the American Statistical Association*, 118(541):242–256.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Inoue, K., Athey, S., and Tsugawa, Y. (2023). Machine-learning-based high-benefit approach versus conventional high-risk approach in blood pressure management. *International Journal of Epidemiology*, page dyad037.

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

Jöbges, S., Vinay, R., Luyckx, V. A., and Biller-Andorno, N. (2020). Recommendations on covid-19 triage: international comparison and ethical analysis. *Bioethics*, 34(9):948–959.

Johansson, F. D., Kallus, N., Shalit, U., and Sontag, D. (2018). Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*.

Johansson, F. D., Shalit, U., Kallus, N., and Sontag, D. (2020). Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*.

Kallus, N. (2019). Classifying treatment responders under causal effect monotonicity. In *International Conference on Machine Learning*, pages 3201–3210. PMLR.

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*.

Kent, D. M., Steyerberg, E., and van Klaveren, D. (2018). Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *Bmj*, 363.

Kluge, E.-H. W. (2007). Resource allocation in healthcare: implications of models of medicine as a profession. *Medscape General Medicine*, 9(1):57.

Korhonen, P. and Syrjänen, M. (2004). Resource allocation based on efficiency analysis. *Management Science*, 50(8):1134–1144.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.

Nandy, P., Yu, X., Liu, W., Tu, Y., Basu, K., and Chatterjee, S. (2022). Generalized causal tree for uplift modeling. *arXiv preprint arXiv:2202.02416*.

National Academies of Sciences, E., Medicine, et al. (2020). *Framework for equitable allocation of COVID-19 vaccine*. National Academies Press.

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., and Verbeke, W. (2020). Uplift modeling for preventing student dropout in higher education. *Decision Support Systems*, 134:113320.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pinkerton, S. D., Johnson-Masotti, A. P., Derse, A., and Layde, P. M. (2002). Ethical issues in cost-effectiveness analysis. *Evaluation and program planning*, 25(1):71–83.

Radcliffe, N. (2007). Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, pages 14–21.

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

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

Rudin, C. and Schapire, R. E. (2009). Margin-based ranking and an equivalence between adaboost and rankboost.

Rzepakowski, P. and Jaroszewicz, S. (2012). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32:303–327.

Sarvet, A. L., Wanis, K. N., Young, J., Hernandez-Alejandro, R., Hernán, M. A., and Stensrud, M. J. (2020). Causal inference with limited resources: proportionally-representative interventions. *arXiv preprint arXiv:2002.11846*.

Schwab, P., Linhardt, L., and Karlen, W. (2018). Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656.*

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

Shalit, U., Johansson, F. D., and Sontag, D. (2017). 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.

Shi, Y., Larson, M., and Hanjalic, A. (2010). Listwise learning to rank with matrix factorization for collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 269–272.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472.

Steck, H., Krishnapuram, B., Dehing-Oberije, C., Lambin, P., and Raykar, V. C. (2007). On ranking in survival analysis: Bounds on the concordance index. *Advances in neural information processing systems*, 20.

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

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

Wang, Y., Wang, L., Li, Y., He, D., Chen, W., and Liu, T.-Y. (2013). A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, volume 8, page 6.

Xia, F., Liu, T.-Y., Wang, J., Zhang, W., and Li, H. (2008). Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966.*

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. (2020). A survey on causal inference. *arXiv preprint arXiv:2002.02770.*

Yip, J. Y.-C. (2021). 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.

Zhang, Y., Bellot, A., and van der Schaar, M. (2020). Learning overlapping representations for the estimation of individualized treatment effects. *arXiv preprint arXiv:2001.04754.*

Zhao, Y., Fang, X., and Simchi-Levi, D. (2017). Uplift modeling with multiple treatments and general response types. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 588–596. SIAM.

Zhou, H., Li, S., Jiang, G., Zheng, J., and Wang, D. (2023). Direct heterogeneous causal learning for resource allocation problems in marketing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5446–5454.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. No
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Yes
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Appendix

A RELATED WORK

CATE Estimation. In recent years, there has been increased interest in estimating the heterogeneous effects of treatments from confounded observational data (Yao et al., 2020). A majority of past works have proposed solutions for overcoming the issue of confounding. Past work has considered learning balanced representations (Shalit et al., 2017; Johansson et al., 2018, 2020; Hassanpour and Greiner, 2020), reweighting using propensity scores (Hassanpour and Greiner, 2019, 2020; Assaad et al., 2021; Li et al., 2018), and using doubly robust proxies (Kennedy, 2020) across a wide variety of machine learning architectures, namely neural networks (Shalit et al., 2017) and random forests (Wager and Athey, 2018). 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 (Sarvet et al., 2020). 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 (Fernández-Loría and Provost, 2022; Athey and Wager, 2021; Kallus, 2019). This path represents a step towards bridging the gap between theory and practice for causal inference. In this 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 (Yadlowsky et al., 2021). 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 (Fernández-Loría and Loría, 2022). In this work, we continued with the no hidden confounders assumption and focus 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 (Rzepakowski and Jaroszewicz, 2012; Betlei et al., 2021). Uplift modeling is a common method used particularly in business and marketing problems (Rzepakowski and Jaroszewicz, 2012; Yadlowsky et al., 2021). One approach towards uplift modeling is to estimate pointwise effects of interventions on an individual basis, similar to CATE estimation (Gutierrez and Gérardy, 2017; Nandy et al., 2022). A secondary approach is to optimize for cumulative gain across intervention thresholds, similar to our goal (Zhao et al., 2017; Devriendt et al., 2020). 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 (Zhao et al., 2017). In this 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 (Zhou et al., 2023). 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 optimal treatment allocation 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 optimal 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 (Cao et al., 2007). 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 (Ibrahim and Carman, 2016; Cao et al., 2007; Xia et al., 2008; Shi et al., 2010). 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 (Järvelin and Kekäläinen, 2002; Wang et al., 2013). 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 (Devriendt et al., 2020; Betlei et al., 2021). In this 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 (Ibrahim and Carman, 2016).

B ADDITIONAL PROOFS

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

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 $f : \mathcal{X} \to \mathbb{R}$ 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 $AUTOC_S(g) > AUTOC_S(f)$.*

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.

□

C 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 $T^M_{k,v}$ by splitting current node $M$ by feature $k$ and value $v$
    Calculate proposed value as $\widetilde{AUTOC}_S(T^M_{k,v})$ by traversing sample $S$ through $T^M_{k,v}$
    if proposed value improves over best value then
        Update best value to proposed value
        Update best k,v to be proposed k,v
    end if
end for
return best k and v if they exist
```

D 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

Table 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

	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 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, result in similar performance.

robust proxies as imputed CATEs. 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 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 (Shalit et al., 2017; Schwab et al., 2018; Hassanpour and Greiner, 2019). 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 CATE 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.

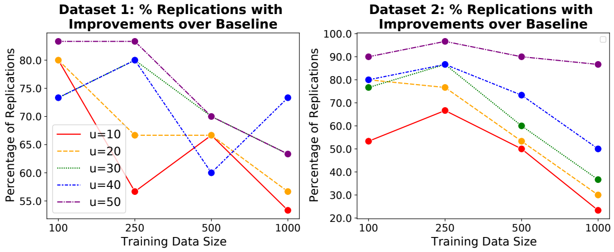

Figure 6: The percentage of replications in which the proposed method outperforms the baseline in terms of ATE^u across different treatment thresholds u and training data size. The proposed method outperforms the baseline in up to $80-90\%$ of replications at different thresholds at low training data size, but the efficacy is only shown at higher treatment thresholds when enough training data is incorporated into the model.

E 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 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 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 (Zhou et al., 2023). 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.

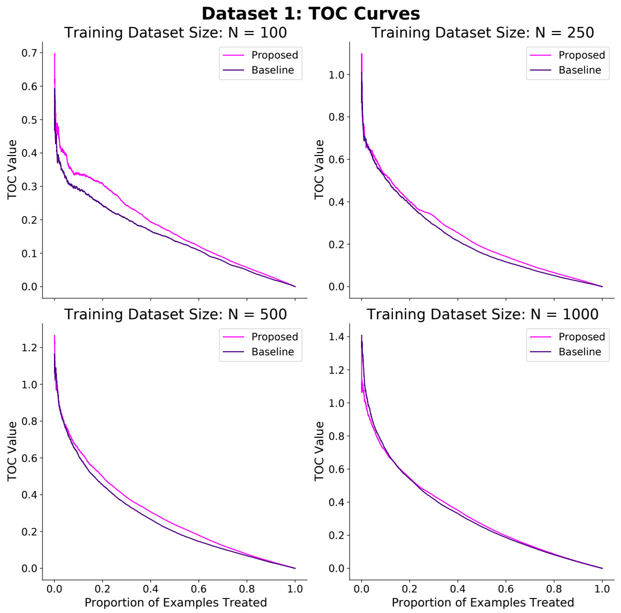

Figure 7: TOC Curves for **Dataset 1**. In low-data settings, our method consistently results in a larger improvement in the ATE of the top percentage of individuals. As more data is included in our model, the improvements of our model are reduced, but our model still results in a larger TOC value across a majority of replications. When all individuals are treated, our method and the proposed method result in no improvement over random.

Results at Specific Treatment Thresholds: To complement the ATE^u results in the main paper, 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 6**. 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, we report the TOC^u values for all $u \in [0, 1]$. These results can be found in **Figure 7** for **Dataset 1**, and **Figure 8** for **Dataset 2**. For

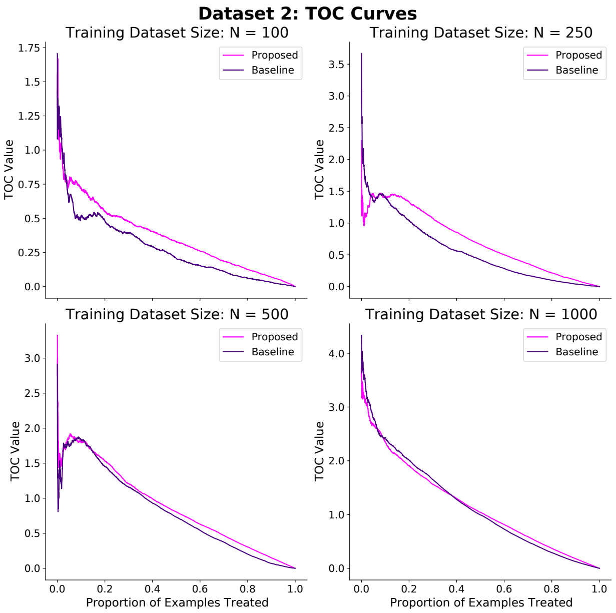

Figure 8: TOC Curves for **Dataset 2**. In low data settings, our method results in a larger improvement in the ATE of the top percentage of individuals, particularly when the treatment threshold is above 10%. When $N = 1000$ data points are used to train the model, the baseline begins to slightly outperform the proposed method, especially at earlier treatment thresholds.

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 9**. 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 better improve upon the baseline in low training data regimes.

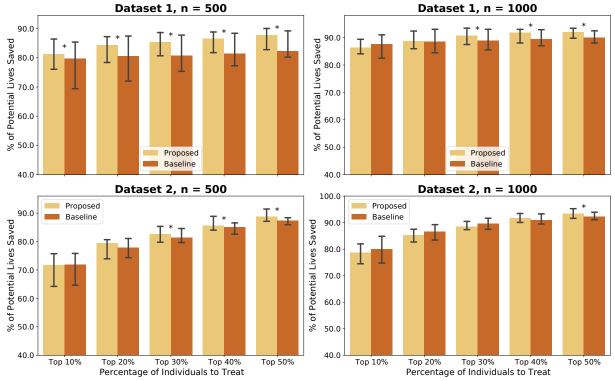

Figure 9: Percentage of potential lives saved compared to the oracle across different treatment settings for high data settings for **Dataset 1** (top) and **Dataset 2** (bottom). Comparisons with asterisks represent scenarios in which the proposed method significantly outperforms the baseline technique 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.