Learning Optimal Dynamic Treatment Regimes Using Causal Tree Methods in Medicine

Theresa Blümlein THERESA@BLUEMLEIN.INFO
ETH Zurich
Zurich, Switzerland

Joel Persson* JPERSSON@ETHZ.CH
ETH Zurich
Zurich, Switzerland

Stefan Feuerriegel FEUERRIEGEL@LMU.DE
LMU Munich
Munich, Germany

Abstract

Dynamic treatment regimes (DTRs) are used in medicine to tailor sequential treatment decisions to patients by considering patient heterogeneity. Common methods for learning optimal DTRs, however, have shortcomings: they are typically based on outcome prediction and not treatment effect estimation, or they use linear models that are restrictive for patient data from modern electronic health records. To address these shortcomings, we develop two novel methods for learning optimal DTRs that effectively handle complex patient data. We call our methods DTR causal trees (DTR-CT) and DTR causal forest (DTR-CF). Our methods are based on a data-driven estimation of heterogeneous treatment effects using causal tree methods, specifically causal trees and causal forests, that learn non-linear relationships, control for time-varying confounding, are doubly robust, and explainable. To the best of our knowledge, our paper is the first that adapts causal tree methods for learning optimal DTRs. We evaluate our proposed methods using synthetic data and then apply them to real-world data from intensive care units. Our methods outperform state-of-the-art baselines in terms of cumulative regret and percentage of optimal decisions by a considerable margin. Our work improves treatment recommendations from electronic health record and is thus of direct relevance for personalized medicine.

1. Introduction

Personalized medicine has emerged as a promising approach for personalizing treatment decisions to patients (Allam et al., 2021; Kosorok and Laber, 2019; Zhao and Zeng, 2013). This is in contrast to the previous "one-size-fits-all" paradigm where patient heterogeneity in treatment responses and risk factors is largely ignored. Personalized medicine offers large benefits, in particular for chronic diseases as these oftentimes require multiple sequential treatments (Chakraborty and Moodie, 2013).

Prior literature has developed methods for learning sequences of individualized treatment decisions. One stream has emerged from statistics and is subsumed under **dynamic**

* Corresponding author

treatment regimes (DTRs) (Chakraborty and Moodie, 2013; Moodie et al., 2007; Tsiatis et al., 2019). DTRs take historical patient information as input and output personalized treatment decisions. Classical DTR methods have relied on linear regression (Chakraborty and Moodie, 2013; Murphy, 2003; Murphy et al., 2001; Tsiatis et al., 2019). These methods benefit from being explainable, yet, because of their simple structure, their use for modern electronic health records is limited. This is the case as such data are typically heterogeneous and embed unknown non-linear relationships between variables. This requires flexible and data-driven models. Another stream of works originates from machine learning, but these methods are primarily black-box (e. g., Liu et al. (2017)) and, in general, are thus not suitable for personalized medicine. Hence, methods are needed for learning optimal DTRs that are explainable yet able to accommodate complex relationships.[1]

In this paper, we develop novel methods, namely DTR-CT and DTR-CF, for learning optimal DTRs. Our proposed methods are based on a data-driven estimation of heterogeneous treatment effects using causal tree methods, specifically causal trees and causal forests. For this purpose, we adapt causal trees (Athey and Imbens, 2016) and causal forests (Athey et al., 2018; Wager and Athey, 2018) from the static setting to the sequential setting in DTRs. As a result, DTR-CT and DTR-CF help in making treatment recommendations from electronic health records. Importantly, our methods learn non-linear relationships, control for time-varying confounding, are doubly robust, and explainable.

We evaluate our methods as follows. (1) We show the effectiveness using synthetic data. Here, our methods outperform baselines in terms of cumulative regret and decision accuracy (i. e., percentage of optimal decisions). Compared to the best performing baseline, DTR-CT and DTR-CF reduce the cumulative regret by more than 20 % in our synthetic data. In addition, DTR-CT and DTR-CF result in optimal treatment decisions at considerably higher rate. (2) We demonstrate the applicability of our proposed methods to real-world medical data from intensive care units. We find that our methods recommend treatment decisions that result in better expected outcomes than under the observed treatment regime.

Our main **contributions** are as follows[2]:

1. To the best of our knowledge, our work is the first to leverage causal trees and causal forests for learning optimal DTRs. Our proposed methods, DTR-CT and DTR-CF, thus learn heterogeneous treatment effects that are subject to non-linear relationships.

2. Our proposed methods are explainable. For example, one can directly visualize the underlying decision tree in DTR-CT. Explainability is a key advantage for clinical relevance in medical decision-making and personalized medicine where the basis of treatment decisions to individual patients must be transparent and intelligible (Amann et al., 2020; Cutillo et al., 2020; Ribeiro et al., 2016).

1. In this paper, we refer to methods that reduce the opaqueness of a DTR as "explainable". We are aware that some authors distinguish "interpretable" (when the decision logic itself is transparent) vs. "explainable" (when insights are based on a surrogate model), see Amann et al. (2020) and references therein for an overview. In this terminology, our DTR-CT would be classified as interpretable, while DTR-CF as explainable.
2. Our codes and results are publicly available via https://github.com/tbluemlein/DTR

3. We outperform state-of-the-art baselines for learning optimal DTRs and demonstrate the applicability of our proposed methods to real-world medical data.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work also provides the following generalizable insights about machine learning for healthcare:

1. Machine learning methods are increasingly developed that are both of high predictive accuracy and explainable (Rudin, 2019). However, previous methods typically focus on *predictions*, while we target *decisions*. As such, we foresee important future work and impact in clinical practice through the use of prescriptive analytics for decision-making.

2. We find that, for personalizing treatment decisions, classical statistical methods and prediction-focused machine learning methods may be sub-optimal. Instead, we see benefits from using machine learning methods that are explicitly developed for causal effect estimation.

3. We empirically demonstrate that our proposed methods can improve on the treatment decisions of medical practitioners. This suggests that practitioners may benefit from adopting the methods in clinical practice. At the same time, this also points to untapped potential for further optimization of clinical decision-making in intensive care units.

2. Related Work

Heterogeneous Treatment Effects. The main goal of personalized medicine is to tailor treatments to patients by considering their heterogeneity. Doing so requires learning the treatment effects heterogeneity across patients. To quantify variation in treatment effects, methods to estimate heterogeneous treatment effects (HTEs) have recently gained in popularity. Examples include causal trees (Athey and Imbens, 2016) and causal forests (Athey et al., 2018; Wager and Athey, 2018), which are designed for estimating HTEs in static settings but have not yet been adapted to learn optimal DTRs.

Dynamic Treatment Regimes. Research on DTRs dates back to Robins (1997, 2004). Historically, these methods relied on linear low-dimensional parametric models estimated with least squares (Moodie et al., 2012; Murphy, 2003; Murphy et al., 2001; Tsiatis et al., 2019; Wallace et al., 2017). Although the statistical properties of these methods are well-known, their simplicity may not be suited for modern electronic health record data that embed unknown non-linearities and complex relationships (Liu et al., 2017).

Machine Learning for DTRs. Recently, machine learning methods have been proposed for accommodating the complexities in learning DTRs from electronic health record data. Although tree-based methods have been proposed for providing explainable treatment recommendations (Speth and Wang, 2021; Speth et al., 2020; Sun and Wang, 2021; Tao et al., 2018; Zhang et al., 2018; Zhao et al., 2015), they have been constructed as CART or random forests and thus do not explicitly model causal effects. However, the latter is a key property for reliably learning HTEs and, thus, to optimize treatment recommendations.

Compared to previous tree-based methods, our proposed methods directly estimate HTEs with a causal interpretation via a tailored splitting criterion for the decision trees. As an alternative to DTRs, other works (Cai et al., 2021; Yauney and Shah, 2018; Wang et al., 2018; Prasad et al., 2017; Raghu et al., 2017) have used reinforcement learning. However, these methods are typically based on variants of deep learning or neural networks (see Yu et al. (2021); Liu et al. (2020) for surveys) and, therefore, are not readily explainable. Finally, some works (Luckett et al., 2019; Nie et al., 2021) have developed data-driven methods for learning optimal DTRs but for different settings to our work. In particular, Luckett et al. (2019) propose the use of v-learning for estimating DTRs based on randomized decision rules in an online setting, and Nie et al. (2021) develop DTRs for when to start and stop treatment.

Research Gap. To the best of our knowledge, no previous study has used causal methods such as causal trees or causal forests to learn optimal DTRs. By leveraging these methods, our methods are able to better learn individualized sequential treatment decisions from modern electronic health records data that are explainable to clinical practitioners.

3. Setup

3.1. Causal Structure of the Data

Let $i = 1, \ldots, N$ refer to patients, and $t = 1, \ldots, T$ refer to time steps. Each patient has m baseline covariates $\mathbf{C}_i = [C_{i1}, \ldots, C_{im}]$ (e. g., risk factors such as age or gender) measured prior to initial treatment, and has p time-varying covariates $\mathbf{X}_{it} = [X_{it1}, \ldots, X_{itp}]$ (e.g., vital sign measurements such as blood pressure or oxygen saturation) measured at every time step $t = 1, \ldots, T$. Given the baseline covariates and time-varying covariates, a binary treatment (or treatment decision) $A_t^{(i)} = A_t^{(i)}(\mathbf{C}^{(i)}, \mathbf{X}_t^{(i)}) \in \{0, 1\}$ is prescribed. Here, $A_{it} = 1$ means that patient i is treated at time step t, and $A_{it} = 0$ that the patient is not treated. In the following, we will omit the patient superscript (i) to simplify notation.

For brevity, let $A_{1:t} = (A_1, \ldots, A_t)$ denote the history of binary treatments up to and including time t, and let $\mathbf{X}_{1:t} = (\mathbf{X}_1, \ldots, \mathbf{X}_t)$ denote the time-varying covariates up to and including time t. Then, $\mathbf{H}_t := (\mathbf{C}, A_{1:t-1}, \mathbf{X}_{1:t}) \in \mathcal{H}_T$ is the complete patient history through time t. The outcome of interest is $Y = Y(\mathbf{H}_T, A_T)$, which is observed at time step $T + 1$ and depends on the full history (w.l.o.g. larger values are preferred). Fig. 1 shows a directed acyclic graph of the causal structure of data.

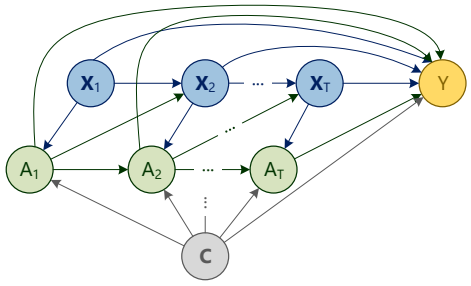


Figure 1: Causal structure of the sequential decision-making problem.

3.2. Target Estimand: Optimal DTR

Our estimands of interest are optimal DTRs, meaning rules for optimizing sequential treatment decisions. Let $\delta \coloneqq (\delta_1(\mathbf{h}_1), \ldots, \delta_T(\mathbf{h}_T)) \in \mathcal{D}$ denote an arbitrary DTR. Here, each $\delta_t \colon \mathcal{H}_t \to \{0, 1\}$ is a mapping from the set of patient histories to the set of treatment alternatives at time t, and $\mathcal{D}$ is the set of all possible DTRs defined by a class of functions used to learn a DTR. Thus, a DTR δ is a sequence of decision rules that personalizes treatments to patients according to their heterogeneous histories.

Following standard practice in the causal inference literature, we use the potential outcomes framework (Neyman, 1923; Rubin, 1974) to express our target estimand. Let $Y(\delta)$ denote the potential outcome that would be realized if treatments were assigned according to a DTR $\delta \in \mathcal{D}$. Moreover, let $V(\delta) \coloneqq \mathbb{E}[Y(\delta)]$ be the so-called *value* of δ, i. e., the mean outcome in the population if every patient was assigned treatments according to δ. The target estimand is the optimal DTR δ^{opt}, defined as

$$\delta^{\text{opt}} \coloneqq \arg\max_{\delta \in \mathcal{D}} V(\delta). \tag{1}$$

3.3. Causal Inference for DTRs

To optimize sequential decisions, we must account for both prognostic and delayed effects of treatments as well as which decisions will be taken in the future. For learning optimal DTRs, Robins (2004) thus defined the so-called *blip* function (from here on also called HTE) as the expected difference in potential outcomes at time T would a patient be treated at time $t \leq T$ or not, conditional on the history and given subsequent optimal decisions. The blip function is thus given by

$$\tau_t(\mathbf{h}_t) \coloneqq \mathbb{E}[Y(1, \delta^{\text{opt}}_{t+1:T}) - Y(0, \delta^{\text{opt}}_{t+1:T}) \mid \mathbf{H}_t = \mathbf{h}_t]. \tag{2}$$

If the HTE was known, the optimal decision at time t would be to assign treatment to the patients for whom treatment is beneficial, $\delta^{\text{opt}}_t = \mathbb{1}\{\tau_t > 0\}$, assuming no treatment costs or restrictions. However, as we can only observe the factual outcome for the observed treatment regimen, the HTE is a counterfactual quantity. To identify it from data, we impose the following assumptions that are standard in the causal inference literature (Robins and Hernán, 2009; Rosenbaum and Rubin, 1983).

Assumption 1 (Consistency): If $A_{1:T} = a_{1:T}$, then $Y(a_{1:T}) = Y$.

Assumption 2 (Stable Unit Treatment Value Assumption): The potential outcome of any patient is independent of the treatment assignment of any other patient.

Assumption 3 (Positivity): If $\mathbb{P}(\mathbf{H}_t = \mathbf{h}_t) \neq 0$, then $\mathbb{P}(A_t = a_t \mid \mathbf{H}_t = \mathbf{h}_t) > 0 \ \forall$ $a_t \in \{0, 1\}$, $t = 1, \ldots, T$.

Assumption 4 (Sequential Exchangeability): $Y(a_t) \perp A_t \mid \mathbf{H}_t \ \forall \ a_t \in \{0, 1\}$, $t = 1, \ldots, T$.

Assumption 1 links the potential outcomes to the observed outcomes by requiring that the two are equal under the same treatment assignments. Assumption 2 implies that there is no interference between treatment assignment and outcomes across patients. Assumption 3 states that, for any possible patient history, treatment assignment is not deterministic. Assumption 4 implies that conditioning on the patient histories is sufficient to remove confounding bias in estimated HTEs.

4. Methods

4.1. Overview

We now introduce our algorithms DTR-CT and DTR-CF. Therein, we combine a (i) backward recursive estimation algorithm and (ii) sequential HTE estimation. For the latter, we use a causal tree for DTR-CT and a causal forest for DTR-CF. In the following, we first introduce our backward recursive estimation algorithm (Section 4.2), which we then explain for our methods DTR-CT (Section 4.3) and DTR-CF (Section 4.4).

The rationale for using a backward recursive algorithm is the following. Backward recursion is a dynamic programming method used for learning optimal DTRs (Murphy, 2003) that addresses two fundamental challenges in sequential decision-making: (1) We must account for both prognostic and delayed effects of past decisions and the current decision (Jiang et al., 2022). (2) Time-varying covariates can be both confounders and mediators of treatment effects depending on their time of measurement in relation to the treatment decision. Hence, conditioning on covariates measured after the treatment in a time step induces a post-treatment bias by "explaining-away" the treatment effect (Chakraborty and Moodie, 2013, Ch. 3.4.2). Thus, fitting a single model to the complete patient history leads to biased DTRs.

Backward recursion solves the above-mentioned challenges by decomposing the sequential decision-problems into a set of single-step problems. Each single-step problem involves fitting a so-called outcome model to estimate the HTE and then solving for the optimal decision at that time step. Because the outcome is only observed at the final time step, we rely on optimizing so-called *pseudo-outcomes* at each time step: the potential outcomes would the patient be treated optimally at later time steps (Chakraborty and Moodie, 2013).

4.2. Backward Recursive Estimation Algorithm

Algorithm 1 presents our estimation procedure for our proposed methods. Algorithm 1 takes patient histories and outcomes $(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T, y)$ as input, and outputs an estimated optimal DTR from the set $\mathcal{D}$ of all possible DTRs defined by the chosen method (i. e., DTR-CT or DTR-CF). Depending on the method chosen, we fit a causal tree or causal forest, respectively, to estimate the HTE per time step. We do so by calling the functions DynamicWeightedCausalTree and DynamicWeightedCausalForest, which are causal trees and causal forests that we adapt from the static setting to the sequential setting for use in the backward recursive algorithm. Both are described in the subsequent sections.

Algorithm 1 starts by initializing a DTR δ as the observed treatment sequence in the data, $\{a_1, a_2, \ldots, a_T\}$, and assigns the pseudo-outcome y' as the observed final time step

Algorithm 1: DTR-CT and DTR-CF

Data: patient histories and outcomes $\{(\mathbf{h}_T, y)\}$
Result: $\hat{\delta}^{\text{opt}} = (\hat{\delta}_1^{\text{opt}}, \ldots, \hat{\delta}_T^{\text{opt}}) \in \mathcal{D}$
Initialize DTR as the observed treatment sequence in the data: $\delta \leftarrow (a_1, \ldots, a_T)$.
Initialize the pseudo-outcomes as the observed outcomes in the data: $Y'_T \leftarrow Y$.
for $t = T, \ldots, 1$ **do**
  Estimate propensity scores: $\hat{\pi}(\mathbf{h}_t) = (1 + \exp(-\mathbf{h}_t^\top \hat{\gamma}_t)^{-1}$.
  if DTR-CT **then**
    $\hat{\tau}_t^{(\ell)}(\mathbf{h}_t) \leftarrow \textsc{DynamicWeightedCausalTree}(y \sim (\mathbf{h}_t, a_t), \text{weight} = (\hat{\pi}(\mathbf{h}_t))^{-1})$
  else if DTR-CF **then**
    $\hat{\tau}(\mathbf{h}_t) \leftarrow \textsc{DynamicWeightedCausalForest}(y \sim (\mathbf{h}_t, a_t), \text{weight} = (\hat{\pi}(\mathbf{h}_t))^{-1})$
  Estimate the optimal treatment decisions:

$$\hat{\delta}^{\text{opt}}(\mathbf{h}_t) \leftarrow \mathbb{1}\{\hat{\tau}(\mathbf{h}_t) > 0\}.$$

  Update the DTR:

$$\hat{\delta} \leftarrow (a_1, \ldots, a_{t-1}, \hat{\delta}^{\text{opt}}(\mathbf{h}_t), \ldots, \hat{\delta}^{\text{opt}}(\mathbf{h}_T));$$

  Update the pseudo-outcome via

$$Y'_{t-1} \leftarrow Y'_t + (\hat{\delta}^{\text{opt}}(\mathbf{h}_t) - a_t)\hat{\tau}(\mathbf{h}_t).$$

end
Assign the estimated optimal decisions to the DTR: $\hat{\delta}^{\text{opt}} \leftarrow \hat{\delta}$
return the estimated optimal DTR $\hat{\delta}^{\text{opt}}$

outcome y for each individual. The recursion is then perform backwards, starting from the last time step and proceeding until the first, to update the DTR and pseudo-outcomes.

For time step $t = T, \ldots, 1$, the backward recursion proceed as follows:

1. We estimate the so-called propensity scores (Rosenbaum and Rubin, 1983) at time t. In our setting with binary treatments, the propensity score is defined as the conditional probability of treatment at time t given the history, i. e., $\pi(\mathbf{h}_t) \coloneqq \mathbb{P}(A_t = 1 \mid \mathbf{H}_t = \mathbf{h}_t)$. We estimate these using logistic regression of A_t on $\mathbf{H}_t$ by solving $(1 + \exp(-\mathbf{h}_t^\top \hat{\gamma}_t)^{-1}$ for maximum likelihood parameter estimates $\hat{\gamma}_t$. Nevertheless, in principle, any probabilistic classifier can be used.

2. Depending on the chosen method (i. e., DTR-CT or DTR-CF), we estimate the HTE by fitting a causal tree or a causal forest to the observed outcomes Y using the history $\mathbf{h}_t$. We follow the DWOLS method (Wallace et al., 2017) and weight the outcomes with the inverse of the estimated propensity scores. See Sections 4.3 (for DTR-CT) and 4.4 (for DTR-CF) for details.

3. We map the estimated HTE $\hat{\tau}(\mathbf{h}_t)$ to an optimal treatment decision $\hat{\delta}^{\text{opt}}(\mathbf{h}_t)$ by assigning the treatment if the HTE is positive, and assigning the control otherwise.

4. By the plug-in principle, we update the DTR by imputing the estimated optimal decision at time t for each individual.

5. We update the pseudo-outcome for the next time step in the backward recursion. Here, the pseudo-outcome at the next iteration $t - 1$ is updated as the pseudo-

outcome at time t, plus the HTE at time t if the observed and estimated optimal treatment decisions differ. Thus, the pseudo-outcome captures the increments (i. e., improvements) in the outcome under the estimated optimal treatment decisions.

Having completed the backward recursion, the algorithm outputs the estimated optimal DTR.

4.3. DTR-CT

We adapt causal trees (Athey and Imbens, 2016) to the backward recursive algorithm for estimating optimal DTRs. Using a splitting criterion at every time step in the backward recursion, we create a partition of the patient population into subpopulations (i. e., a tree Π_t consisting of leaves $\ell \in \mathcal{L}_t$, where $\mathcal{L}_t$ is a finite set of all leaves in time step t). The treatment effects are then estimated per heterogeneous leaf to obtain HTEs.

To ensure unbiasedness and avoid overfitting, we use the so-called *honest* splitting criterion (Athey and Imbens, 2016) for the partitioning that decouples model selection from model estimation. It does so by training the causal tree on a training set and estimating the HTEs with the trained causal trees on an estimation set.[3] The splitting criterion is then given by the expected mean squared error (EMSE) of the HTEs over the training set and the estimation set.

Let $\Pi_t = \{l_1, \ldots, l_{|\mathcal{L}_t|}\}$ denote a partitioning of histories $\{\mathbf{h}_t\}$ into $|\mathcal{L}_t|$ leaves such that $\bigcup_{\mathbf{\Pi}_t} \ell = \mathcal{L}_t$. Moreover, let $\mathcal{S}_t^{\text{tr}}$ and $\mathcal{S}_t^{\text{est}}$ denote the training set and estimation set, respectively, such that $\mathcal{S}_t^{\text{tr}} \cap \mathcal{S}_t^{\text{est}} = \emptyset$. Adapting the EMSE splitting criterion from the static setting to the sequential setting and for use with our backward recursive algorithm, we seek to find a partitioning that maximizes

$$-\operatorname{EMSE}_{\tau_t(\mathbf{h}_t)}(\Pi_t) = \mathbb{E}_{\mathcal{H}_t}[\tau^2(\mathbf{H}_t; \Pi_t)] - \mathbb{E}_{\mathcal{S}_t^{\text{est}}, \mathcal{H}_t}[\mathbb{V}[\tau^2(\mathbf{H}_t; \mathcal{S}_t^{\text{est}}, \Pi_t)]]. \tag{3}$$

We further adapt its cross-validated sample estimator to our setting, given by

$$\begin{aligned} -\widehat{\operatorname{EMSE}}_{\tau_t(\mathbf{h}_t)}(\mathcal{S}_t^{\text{tr,cv}}, N_t^{\text{est}}, \Pi_t) \coloneqq & \frac{1}{N_t^{\text{tr,cv}}} \sum_{i \in \mathcal{S}_t^{\text{tr,cv}}} \hat{\tau}^2(\mathcal{H}_t; \mathcal{S}_t^{\text{tr,cv}}, \Pi_t) - \left(\frac{1}{|\mathcal{S}_t^{\text{tr,cv}}|} + \frac{1}{|\mathcal{S}_t^{\text{est}}|} \right) \cdot \\ & \sum_{l \in \Pi_t} \left(\frac{S^2_{\mathcal{S}^{\text{tr,cv}}_{(A_t=1)}}(l)}{|\mathcal{S}^{\text{tr,cv}}_{(A_t=1)}|/N_t} + \frac{S^2_{\mathcal{S}^{\text{tr,cv}}_{(A_t=0)}}(l)}{1 - |\mathcal{S}^{\text{tr,cv}}_{(A_t=0)}|/N_t} \right), \end{aligned} \tag{4}$$

where $|\mathcal{S}_t^{\text{tr,cv}}|$ and $|\mathcal{S}_t^{\text{est}}|$ are the number of observations for model selection and model estimation, $|\mathcal{S}^{\text{tr,cv}}_{(A_t=1)}|$ and $|\mathcal{S}^{\text{tr,cv}}_{(A_t=0)}|$ the number of treated and control training observations, and N_t the total number of observations until time t. Finally, $S^2_{\mathcal{S}^{\text{tr,cv}}_{(A_t=1)}}(l)$ and $S^2_{\mathcal{S}^{\text{tr,cv}}_{(A_t=0)}}(l)$ is the variance of the treated and control outcomes in leaf l. The superscript (cv) denotes the cross-validation fold. Intuitively, this splitting criterion rewards partitions with homogeneity within leaves and heterogeneity across leaves in terms of their treatment effects. See Athey and Imbens (2016) for details.

3. We use the terms training set for model selection and estimation set for model estimation as in the original paper by Athey and Imbens (2016). In machine learning, the terms training set (for model estimation) and development set (for model selection) are also common.

Having found a partitioning Π_t that maximizes Eq. (4), we estimate the HTE at time step t in each leaf ℓ on the observed outcomes with

$$\hat{\tau}_t^{(\ell)}(\mathbf{h}_t) = \frac{\sum_{\{i:A_{it}=0, \mathbf{h}_{it}\in\ell\}} \hat{\pi}_{it}^{-1} Y_i}{|\{i : A_{it} = 1, \mathbf{h}_{it} \in \ell\}|} - \frac{\sum_{\{i:A_{it}=1, \mathbf{h}_{it}\in\ell\}} \hat{\pi}_{it}^{-1} Y_i}{|\{i : A_{it} = 0, \mathbf{h}_{it} \in \ell\}|}. \tag{5}$$

Here, $\mathbf{h}_{it} \in \ell$ denotes that patient i with history $\mathbf{h}_{it}$ falls in leaf ℓ and $\hat{\pi}_{it}^{-1} = (\hat{\pi}(\mathbf{h}_{it}))^{-1}$ is the inverse of the estimated propensity score. In every leaf ℓ, there are both treated and untreated patients. Hence, the time-specific HTE is estimated leaf-wise as the difference in average outcomes between treated and untreated patients (Athey and Wager, 2019).

Algorithm 2 states the function DynamicWeightedCausalTree that we call inside the backward recursive algorithm. Both together yield our DTR-CT.

Algorithm 2: DynamicWeightedCausalTree

Data: patient histories, observed outcomes, and estimated propensity scores until time step t: $(\mathbf{h}_t, y, \hat{\pi}_t)$
Result: estimated HTE at time step t, i. e. $\hat{\tau}_t^{(\ell)}$, for all $\ell \in \mathcal{L}_t$
Split the data into a training set $\mathcal{S}_t^{\text{tr}}$ and an estimation set $\mathcal{S}_t^{\text{est}}$ (for honesty)
Find partitioning $\mathbf{\Pi}_t$ that maximizes Eq. (4) on $\mathcal{S}_t^{\text{tr}}$
return partition $\mathbf{\Pi}_t$ of the training set into leaves $\mathcal{L}_t$
for *every leaf* $\ell \in \mathcal{L}_t$ **do**
  Estimate the HTE at time step t, i. e., $\hat{\tau}_t^{(\ell)}(\mathbf{h}_t)$, on estimation set $\mathcal{S}_t^{\text{est}}$ using Eq. (5)
end
return HTE at time step t, $\hat{\tau}^{(\ell)}(\mathbf{h}_t)$, for all $\ell \in \mathcal{L}_t$

4.4. DTR-CF

We now tailor causal forests to the sequential setting for with the backward recursive algorithm. A causal forest is an ensemble of causal trees (Athey et al., 2018; Wager and Athey, 2018), serving as an adaptive nearest-neighbor estimator by combining estimates from many causal trees. However, the HTE for a given patient is not estimated as a simple average of HTEs but as weighted average over the local patient neighborhood. Adapting the causal forests HTE estimator from static settings to use in our backward recursive algorithm, we get

$$\hat{\tau}(\mathbf{h}_{it}) = \frac{\sum_{j=1}^{n} \alpha_j(\mathbf{h}_{jt})(Y_j - \hat{Y}^{(-j)})(A_{jt} - \hat{\pi}_t^{(-j)})}{\sum_{j=1}^{n} \alpha_j(\mathbf{h}_{jt})(A_{jt} - \hat{\pi}_t^{(-j)})^2}. \tag{6}$$

Here, Y_j is the observed outcome for patient j, and $\hat{Y}^{(-j)} := \hat{\mathbb{E}}[Y^{(-j)} \mid \mathbf{H}_{jt} = \mathbf{h}_{jt}]$ is its predicted value. We follow Athey et al. (2018) and Athey and Wager (2019) and incorporate the propensity score estimate $\hat{\pi}_t^{(-j)}$ via the differences $(A_{jt} - \hat{\pi}_t^{(-j)})$. Essentially, this corrects for lack of external validity due to unequal treatment assignment probabilities among patients. We also apply the honesty principle of predicting the outcome and propensity score "out-of-bag" not using data of patient j; hence, the subscript $(-j)$. Moreover,

$$\alpha_j(\mathbf{h}_{jt}) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{1}\{\mathbf{h}_{jt} \in \ell^{(b)}(\mathbf{h}_{jt})\}}{\mid \ell^{(b)}(\mathbf{h}_{it}) \mid} \tag{7}$$

is a data-adaptive kernel weight that reflects closeness in terms of covariate values, i. e., the frequency a patient $j \neq i$ falls into the same leaf $\ell^{(b)}$ as the patient of interest i. To construct the neighborhood for each single patient i, we calculate weights for every other patient $j \neq i$. See Athey and Wager (2019) for details on Eq. (6) in the static setting.

Algorithm 3 states the function DynamicWeightedCausalForest that we call inside the backward recursive algorithm. Both together yield our DTR-CF.

Algorithm 3: DynamicWeightedCausalForest

Data: patient histories, observed outcomes, and estimated propensity scores until time step t: $(\mathbf{h}_t, y, \hat{\pi}_t)$
Result: estimated HTEs at time step t, i. e., $\hat{\tau}(\mathbf{h}_t)$
for *every tree* $b = 1, \ldots, B$ **do**
  Split the data into a training set $\mathcal{S}_t^{\text{tr},b}$ and an estimation set $\mathcal{S}_t^{\text{est},b}$ (for honesty)
  Find partitioning $\mathbf{\Pi}_t^{(b)}$ that maximizes Eq. (4) on $\mathcal{S}_t^{\text{tr},b}$
  return partition $\mathbf{\Pi}_t^{(b)}$ of the training set into leaves $\mathcal{L}_t^{(b)}$
  for *every leaf* $\ell \in \mathcal{L}_t^{(b)}$ **do**
    Estimate HTE at time step t, i. e., $\hat{\tau}_t^{(\ell,b)}(\mathbf{h}_t)$, on estimation set $\mathcal{S}_t^{\text{est,b}}$ using Eq. (5)
  end
  return HTE at time step t, $\hat{\tau}^{(\ell)}(\mathbf{h}_t)$, for all $\ell \in \mathcal{L}_t$
end
Calculate weights $\alpha_i(\mathbf{h}_t)$ using Eq. (7)
Calculate HTE at time step t, i. e., $\hat{\tau}_t(\mathbf{h}_t)$, as weighted average of $\{\hat{\tau}_t^{(\ell,b)}(\mathbf{h}_t)\}_{b=1}^{B}$ using Eq. (6)
return HTE at time step t, i. e., $\hat{\tau}_t(\mathbf{h}_t)$

4.5. Double Robustness

In our methods, we weight the observed outcomes with the inverse propensity score estimates when estimating the HTEs using causal trees or causal forests. This approach is analogous to the DWOLS method (Wallace et al., 2017) for learning optimal DTRs. With DWOLS, the HTEs are modeled as linear parametric regression functions estimated with weighted ordinary least squares, where the weights are given by the inverse of the estimated propensity scores. Wallace et al. (2017) shows that this weighting procedure results in double robustness. Here, double robustness refers to that, per time step, only one of the propensity score model or HTE model needs to be correctly specified with unbiased estimators to get unbiased estimates of the HTE, and thus, the optimal decisions. In contrast to DWOLS, our proposed methods estimate the HTEs using causal trees and causal forests. As both of these unbiased estimators of HTEs (Athey et al., 2018; Wager and Athey, 2018; Athey and Imbens, 2016), and maximum likelihood is an unbiased estimator of logistic regression parameters, it thus follows that our proposed methods are also doubly robust.

4.6. Explainability

Both DTR-CT and DTR-CF are explainable:

- DTR-CT: At every time step, we can visualize the decision tree of which variables explain variation in the HTE across patients, and can thus be inspected by medical practitioners when deciding on treatments.

- DTR-CF: At every time step, one can inspect the variable importance in the outcome model estimated with the causal forest, and thus, which variables are most prescriptive for treatment decision-making.

5. Simulation Studies

We perform simulation studies using synthetic data to demonstrate the effectiveness of our methods. Specifically, we compare the estimated optimal DTRs against a ground-truth (i. e., oracle) optimal DTR.

5.1. Data Generating Processes

We generate synthetic data analogous to previous works on DTRs (Moodie et al., 2012; Wallace and Moodie, 2015; Wallace et al., 2017). Specifically, we choose a similar number of patients histories, similar lengths for the patient histories, similar covariates. To this end, we simulate $N = 5,000$ patient histories of length $T = 3$. Each patient history consists of two baseline covariates $C_k \sim \mathcal{N}(\mu_k, 1),\ k = 1, 2$, five continuous time-varying covariates generated as $X_{1,j} \sim \mathcal{N}(\mu_j, 1)$ at $t = 1$ and evolving as autoregressive processes of order 1, $X_{t+1,j} \sim \mathcal{N}(X_{t,j} + \beta_j A_t, 2)$, over $t = 2, 3$ for $j = 1, \ldots, 5$, and two categorical time-varying covariates generated as $X_{1,j} \sim \text{Binom}(N, 3, p)$ at $t = 1$ and evolving as AR(1) processes $\mathbf{X}_{t+1,j} \sim \mathbf{X}_{t,j} + A_t \cdot \text{Binom}(N, 3, p)$ for $t = 2, 3$ with p_j being the success probability for $j = 6, 7$. The categorical covariates have four levels in the first time step $t = 1$ and may have more than three levels at steps $1 < t \leq 3$ due to the additive structure over the time steps.

Scenarios. We consider two scenarios with varying complexity: (1) linear interactions and (2) non-linear interactions. Let $\tau_t^{(1)} \coloneqq f_t(\mathbf{H}_t)$ and $\tau_t^{(2)} \coloneqq f_t(g_t(\mathbf{H}_t))$ for $t = 1, \ldots, T$ denote the true treatment effect under scenario 1 and 2, respectively. Here, $f_t(\cdot)$ is a function of arbitrary interactions between (predominantly categorical) patient history covariates, and $g_t(\cdot)$ is a non-linear function of the patient history. Hence, in scenario 1, the true treatment effect is a linear function of patient history interactions, whereas, in scenario 2, it is a non-linear function of their interactions. We construct the functions with a maximum interaction depth of three in both scenarios. We generate the true treatment effects such that the portion of optimally treated patients differs from the portion of actually treated patients. This mimics real-world settings where some patients are sub-optimally treated and, thereby, implies a potential for improvement by an optimal DTR.

The outcome is defined as $Y^{(k)} \coloneqq \sum_{t=1}^{T} \phi_t^{(k)}(\mathbf{C}, \mathbf{X}_t) + A_t \cdot \tau_t^{(k)} + \varepsilon$, where $\phi_t^{(k)}(\cdot)$ is the scenario $k = 1, 2$ treatment-free function that determines the outcome in absence of treatment at time t, $A_t \cdot \tau_t^{(k)}$ is the treatment effect increment per scenario k for the treated, and $\varepsilon \sim \mathcal{N}(0, 1)$ is an error term. We define the true optimal treatment decision for scenario k as $\delta_{tk}^{\text{opt}} = \mathbb{1}\{\tau_t^{(k)} > 0\}$. The counterfactual outcome obtained under the true optimal DTR is then given by $Y_k^{\text{opt}} = \sum_{t=1}^{T} \phi_t^{(k)}(\mathbf{C}, \mathbf{X}_t) + \delta_{tk}^{\text{opt}} \cdot \tau_t^{(k)}$. For details, see Appendix B and `data_generating_process.R` at https://github.com/tbluemlein/DTR.

5.2. Baselines

We compare our proposed methods to five state-of-the-art baselines for learning optimal DTRs from observational data (see Appendix A for details).

Q-learning (Moodie et al., 2012) estimates the HTEs for the treatment decision rules by modeling the potential outcomes separately as linear functions using OLS, and then taking their difference (Chakraborty and Moodie, 2013, Ch. 3.4.1).

DWOLS (Wallace and Moodie, 2015) extends Q-learning by weighted ordinary least squares with inverse propensity score weights (Wallace and Moodie, 2015). Thus, this baseline is doubly robust.

G-estimation (Robins, 2004) is another doubly robust regression-based estimator of optimal DTRs. It directly models the HTE as the difference in expected potential outcomes using generalized estimating equations (Chakraborty and Moodie, 2013, Ch. 4.3).

CART (Breiman et al., 1984) use a splitting criterion that aggregates homogeneous patients in terms of their outcomes. We fit separate regression trees for the treated and the control groups and then estimate the HTEs as the mean difference in predicted outcomes between the groups. This serves as machine learning baseline to DTR-CT as it is also decision tree but not explicitly developed for the purpose of estimating causal HTEs.

k-NN regression (Beygelzimer et al., 2013) estimates the HTEs by taking the difference between the k-nearest patient in the treated and control group in terms of their history neighbourhood. As non-adaptive neighborhood estimator, and following previous research on causal forests (Athey et al., 2018), it serves as reasonable baseline to DTR-CF that uses adaptive neighborhood matching.

The first three baselines are classical DTR estimators from statistics built on semi-parametric regression functions. In contrast, DTR-CT and DTR-CF are non-parametric and data-driven machine learning methods. The last two baselines thus serve as machine learning baselines. While these are inherently non-parametric and data-driven, they lack the explainability of our proposed methods and the statistical baselines. Explainability is a key advantage for clinical relevance in medical decision-making and personalized medicine where the basis of treatment decisions to individual patients must be transparent and intelligible (Amann et al., 2020; Cutillo et al., 2020; Ribeiro et al., 2016). We thus only compare against baselines that are **explainable** or machine learning baselines that has also been used in previous research (Athey et al., 2018)).

5.3. Implementation Details

We run Algorithm 1 with DTR-CT and DTR-CF. For the baselines, we also use Algorithm 1 but with the baselines instead of DynamicWeightedCausalTree or DynamicWeightedCausalForest to estimate the HTEs per time step, so that these output sequential treatment recommendations. For each method, we estimate the HTEs using the entire patient history $\mathbf{H}_t$ up until the time step in the backward recursion. Thus, all methods have access to all covariates. For Q-learning, DWOLS, and G-estimation, the outcome model is separated into two components (see Eq. (11) in Appendix A.1) that are regressed on $(\mathbf{H}_t, a_t\mathbf{H}_t)$. We use the same sets of covariates to estimate both component by inputting the entire patient history $\mathbf{H}_t$ into each. This mimics real-world applications where one typically controls for all pre-treatment variables in an attempt to satisfy sequential ex-

changability (Assumption 4 in Sec. 3.3). We also estimate the baselines based on CART and k-NN regression as well as our DTR-CT and DTR-CF methods using the entire patient history up to the respective time step. For each method that use the propensity score, we estimate it in the same way by fitting a logistic regression of A_t on $\mathbf{H}_t$ with parameter estimates obtained via maximum likelihood. For additional details on how we implemented the machine learning methods, see Appendix B.3.

5.4. Performance Evaluation

For the synthetic data, we know the oracle (i. e., true) optimal treatment decisions for benchmarking. We use 75 % of the generated data for training and 25 % for testing. Based on the optimal DTRs estimated on the training data, we predict optimal treatment decisions on the test data. We repeat estimation, prediction, and evaluation over 100 runs, each time simulating new data from the data generating process. We compute the cumulative regret and decision accuracy on the test set using the true optimal DTR. For both quantities, we report mean and standard deviation (SD).

Cumulative Regret. The regret at time t of an estimated optimal DTR $\hat{\delta}^{\text{opt}}$, denoted by $\rho_t(\hat{\delta}^{\text{opt}})$, is defined as difference between the (potentially suboptimal) estimated and the true optimal outcome Y, expressed in terms of the true HTE, i. e.,

$$\rho_t(\hat{\delta}^{\text{opt}}) = (\delta_t^{\text{opt}} - \hat{\delta}_t^{\text{opt}}) \cdot \tau_t \geq 0, \quad t = 1, \ldots, T. \tag{8}$$

The regret at any time step t is always non-negative: ρ_t equals zero if the estimated optimal decisions equal the true optimal decisions, and it is positive if they do not. The regret is zero (and thus minimal) if the estimated optimal decisions equals the oracle optimal decisions. Intuitively, the regret thus captures the extent to which estimated incorrect decisions impair the outcome. We obtain the cumulative regret for a given patient by taking the sum of her regrets over all time steps. We then average the patient-specific cumulative regrets over all patients and simulation runs. Smaller values are better.

Decision Accuracy. The decision accuracy describes the proportion of estimated optimal treatment decisions that match the true optimal treatment decisions from the oracle:

$$\text{acc}_t(\hat{\delta}_t^{\text{opt}}) = \frac{\mathbb{1}\{\hat{\delta}_t^{\text{opt}} = \delta_t^{\text{opt}}\}}{N} \in [0, 1], \quad t = 1, \ldots, T. \tag{9}$$

We average the decision accuracy over all time steps, patients, and runs. Larger values are better.

5.5. Results

We evaluate the different methods for learning optimal DTRs with regard to the cumulative regret (Table 1) and decision accuracy (Table 2). Both DTR-CT and DTR-CF outperform the best-performing baseline approach by a large margin. For scenario 1 (linear interactions), DTR-CT and DTR-CF reduce the cumulative regret by 23.0 % and 27.1 %, respectively. For scenario 2 (non-linear interaction), both reduce the cumulative regret by 25.1 % and 48.2 %, respectively. In addition, DTR-CT and DTR-CF produce the true optimal treatment decisions in 7.2 % and 14.4 % more cases in the scenario 2 (non-linear

interactions). We remind the reader that, due to the importance of explainability in medical decision-making, we only compare against baselines that are inherently **explainable** or those that have also been used in previous research on causal tree methods (Athey et al., 2018).

Method	(1) Linear interactions		(2) Non-linear interactions	
	Train	Test	Train	Test
Q-learning (Moodie et al., 2012)	23.746 (0.532)	23.730 (0.739)	41.066 (1.630)	41.487 (3.805)
DWOLS (Wallace and Moodie, 2015)	23.491 (0.509)	23.566 (0.821)	40.756 (1.659)	40.705 (1.990)
G-estimation (Robins, 2004)	55.747 (2.238)	55.807 (2.462)	20.362 (1.976)	20.434 (2.068)
CART (Breiman et al., 1984)	4.096 (0.876)	4.188 (0.956)	25.388 (6.331)	28.338 (26.967)
k-NN Regression (Beygelzimer et al., 2013)	29.510 (0.872)	26.005 (1.149)	24.414 (2.508)	31.750 (26.032)
DTR-CT (proposed)	3.204 (1.107)	3.223 (1.131)	15.200 (8.307)	15.309 (8.471)
DTR-CF (proposed)	**3.035** (0.168)	**3.052** (0.197)	**10.622** (1.496)	**10.582** (1.817)

Reported: Mean (SD). Lower values are better. Best performing approach in bold.

Table 1: Cumulative regret of baselines and proposed methods for train and test set.

Method	(1) Linear interactions		(2) Non-linear interactions	
	Train	Test	Train	Test
Q-Learning (Moodie et al., 2012)	69.5 (0.4)	69.5 (0.7)	56.3 (0.4)	55.9 (0.8)
DWOLS (Wallace and Moodie, 2015)	69.8 (0.4)	69.8 (0.7)	56.2 (0.8)	56.2 (1.0)
G-Estimation (Robins, 2004)	34.9 (2.5)	34.8 (2.7)	56.1 (1.2)	56.0 (1.4)
CART (Breiman et al., 1984)	82.7 (2.1)	82.5 (2.1)	59.0 (3.4)	58.4 (3.4)
k-NN Regression (Beygelzimer et al., 2013)	64.3 (1.0)	69.3 (1.2)	54.7 (1.0)	49.5 (1.7)
DTR-CT (proposed)	83.9 (2.6)	83.9 (2.6)	65.8 (5.3)	65.6 (5.4)
DTR-CF (proposed)	**82.6** (1.5)	**82.6** (1.6)	**72.7** (0.8)	**72.8** (1.0)

Reported: Mean (SD). Higher values are better. Best performing approach in bold.

Table 2: Decision accuracy [%] of baselines and proposed methods for train and test set.

Sensitivity analysis. The performance gains of our methods are robust to changes in hyperparameters. Specifically, we vary the minimum number of observations of treated and untreated patients per leaf between values of 10 and 70. We find little change in cumulative regret and decision accuracy, and that DTR-CF outperforms DTR-CT overall.

6. Empirical Application to Real-World Medical Data

6.1. Setup

We demonstrate the applicability of our proposed methods to real-world observational data of patients admitted to intensive care units (ICUs). Here, our aim is to sequentially prescribe mechanical ventilation or vasopressin such that length of stay is reduced.

We use the MIMIC-III dataset and follow standard preprocessing (Hatt and Feuerriegel, 2021; Johnson et al., 2016; Kuzmanovic et al., 2021; Oezyurt et al., 2021; Wang et al.,

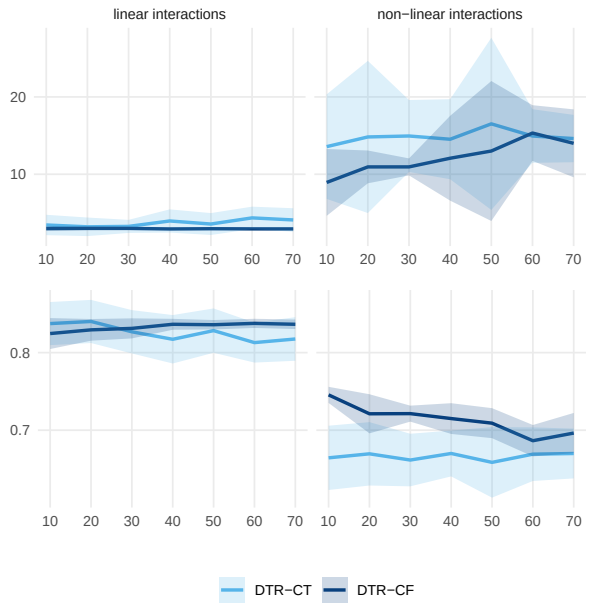


Figure 2: Cumulative regret (y-axis on top panels) and decision accuracy (y-axis on bottom panels) against the minimum number of treated and untreated patients per leaf (x-axis) of DTR-CT and DTR-CF in simulation studies with DGPs featuring linear interactions or non-linear interactions. The performance is evaluated on the test data over 100 runs, each time with new data generated from the respective DGP. The point estimates are the mean across the runs. The shaded regions show the standard deviation across the runs.

2020). We consider a smaller time of the stay in the ICU as proxy for patient health to be maximized. We control for seven time-varying covariates with vital signs: blood pressure [mean, diastolic, systolic], heart rate, oxygen saturation, respiratory rate, and temperature. We further control for six baseline covariates that may confound treatment effects: age, ethnicity, gender, race, insurance, admission type, and first care unit (i.e., the hospital unit at which the patient was initially treated). As treatments, we consider mechanical ventilation and vasopressin. Both are key measures for life support, but also come with side-effects that oftentimes prolong the stay.

We examine the patients who have been treated at least ten times, resulting in $N = 8,059$ patients with histories over the first $T = 10$ hours at the ICU. We imputed missing values of all time-varying covariates except temperature by their medians. We imputed missing values of temperature by a linear model using all other time-varying covariates as predictors as it exhibits the highest percentage of missing data amongst them. For DTR-CT and DTR-CF, we estimate the models using histories over three time steps to detect treatment effects delayed up to 3 time steps.

6.2. Results

For real-world observational data, the true optimal outcome and the true optimal decisions are never known. Hence, we cannot evaluate the performance of our methods against an

oracle DTR, as we did in Sec. 5. Instead, we evaluate our methods by their ability to recommend treatment decisions that result in greater outcomes than those observed in the real-world medical data. The evaluation metric is given by

$$r_t(\hat{\delta}_t^{\text{opt}}) = (\hat{\delta}_t^{\text{opt}} - a_t)\hat{\tau}_t, \quad t = 1, \ldots, T. \tag{10}$$

The metric captures the expected gain in outcome to a given patient if they would be treated according to an estimated optimal DTR learned by one of our proposed methods instead of her factual treatment regimen. A larger value of the metric implies the treatment decisions of the estimated optimal DTR differs from the observed decision in the data and thus indicate an improvement.

Approach	**Treatment**	
	Ventilation	Vasopressin
DTR-CT (proposed)	2.176	2.097
DTR-CF (proposed)	0.233	0.203

Cumulative expected rewards over time, averaged over patients.

Table 3: Cumulative expected rewards for proposed methods on the MIMIC-III data.

Overall, both DTR-CT and DTR-CF achieve a non-negative cumulative reward (see Table 3). Hence, both yield better estimated outcomes than the observed treatment regime. This suggests that our proposed methods offer improvements in patient care with real-world data.

We refrain from comparisons to other DTR baselines as the estimated HTE is scale-variant and approach-specific. Rather, the purpose of the empirical application is to demonstrate the use and benefits of our methods on real-world data.

6.3. Explainability

Another advantage of our proposed methods is that they support standard tools for explainability.

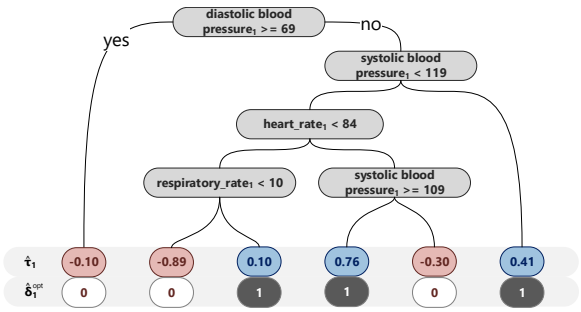


Figure 3: Estimated causal tree within DTR-CT.

Estimated Causal Tree. Fig. 3 shows the estimated causal tree of DTR-CT of a specific time step. Patients within a leaf are homogeneous, but leaves are heterogeneous in

terms of treatment effects. Every leaf shows the estimated treatment effect for its specific subgroup associated with the estimated optimal treatment decision $\hat{\delta}_1^{\text{opt}}$ and whether to apply ventilation. The variables responsible for splitting patients are the variables prescriptive of the recommended treatment decision in that time step. Hence, medical practitioners can compare the causal tree of an estimated optimal DTR against clinical practice and domain knowledge. Even if an estimated optimal DTR would not capture all nuances in clinical practice, it may still be useful for providing personalized decision support.

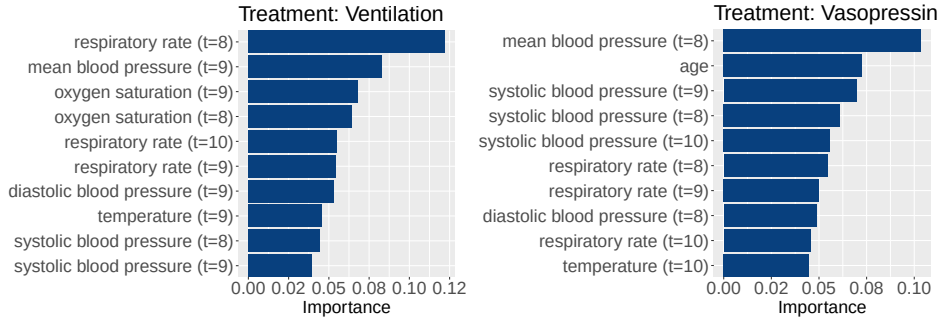


Figure 4: Variable importance within DTR-CF explaining which patient baseline covariates determine treatment assignment (here: $t = 10$).

Variable Importance. Fig. 4 shows the relative importance of the 10 variables that most strongly predict treatment effect heterogeneity in the outcome model estimated with causal forest, and thus, for the estimated optimal treatment decisions by DTR-CF. For illustrative purposes, we only show variable importance measures for a single time step (here: $t = 10$). We find that decisions to apply ventilation are largely based on patients' respiratory rate and oxygen saturation at the two preceding time steps, whereas to prescribe vasopressin, mean and systolic blood pressure over the two preceding time steps are most important. This can be expected from medical knowledge and thus adds to the validity of our results.

7. Discussion

In this paper, we propose two new methods for learning optimal DTRs for sequential treatment decision-making in medicine. Our methods build on causal trees and causal forests that we adapt from estimating static HTEs to estimating optimal DTRs in the sequential setting. The proposed methods flexibly learn non-linear relationships in patient data when personalizing treatment decisions, and account for additional patient heterogeneity by allowing the variables that are prescriptive of the optimal decisions to vary across patients. This is an advantage over classical baselines that assume the relationships to be known and the same variables to be prescriptive for all patients.

Moreover, our methods achieve robust performance across different hyperparameter settings (see Fig. 2), implying little to no need for hyperparameter tuning. Finally, our methods address the need for explainability in medical practice. For instance, medical practitioners may inspect the estimated causal tree to understand which factors explain suggested

treatment decisions and compare those against existing medical domain knowledge. This enables practitioners to better understand and validate optimal DTRs.

Limitations and Future Work Our work is subject to limitations that open interesting paths for future work. First, we evaluate our methods over two time steps. This is consistent with prior works (Moodie et al., 2012; Wallace and Moodie, 2015; Wallace et al., 2017). Future work may investigate longer horizons, in case these are needed for a specific medical condition. Second, we explicitly benchmark our method against baselines that are also explainable. As such, our comparisons include linear models, tree models, and clustering. Explainability is a key requirement for clinical relevance in medical decision-making and personalized medicine where the basis of treatment decisions to individual patients must be transparent and intelligible (Amann et al., 2020; Cutillo et al., 2020; Ribeiro et al., 2016). Third, we acknowledge that we demonstrated the effectiveness of our method through simulation studies. A promising path for future work is to implement our methods in the operational decision-making of healthcare providers to demonstrate the utility for patients in the field.

Acknowledgments

We thank Milan Kuzmanovic for his support in data processing. Joel Persson and Stefan Feuerriegel acknowledge funding from the Swiss National Science Foundation (grant 186932).

References

Ahmed Allam, Stefan Feuerriegel, Michael Rebhan, and Michael Krauthammer. Analyzing patient trajectories with artificial intelligence. *Journal of Medical Internet Research*, 23 (12):e29812, 2021.

Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I Madai. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1):1–9, 2020.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *PNAS*, 113(27):7353–7360, 2016.

Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Observational Studies*, 5(2):37–51, 2019.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2018.

Alina Beygelzimer et al. FNN: Fast nearest neighbor search algorithms and applications. *R package*, 2013.

Matthew Blackwell. A selection bias approach to sensitivity analysis for causal effects. *Political Analysis*, 22(2):169–182, 2014.

Leo Breiman, JH Friedman, R Olshen, and CJ Stone. *Classification and regression trees*. Wadsworth, 1984.

Hengrui Cai, Chengchun Shi, Rui Song, and Wenbin Lu. Deep jump learning for off-policy evaluation in continuous treatment settings. *Advances in Neural Information Processing Systems*, 34, 2021.

Bibhas Chakraborty and EE Moodie. *Statistical methods for dynamic treatment regimes*. Springer, 2013.

Christine M Cutillo, Karlie R Sharma, Luca Foschini, Shinjini Kundu, Maxine Mackintosh, and Kenneth D Mandl. Machine intelligence in healthcare: Perspectives on trustworthiness, explainability, usability, and transparency. *npj Digital Medicine*, 3(1), 2020.

Tobias Hatt and Stefan Feuerriegel. Sequential deconfounding for causal inference with unobserved confounders. *arXiv:2104.09323*, 2021.

Cong Jiang, Michael Wallace, and Mary Thompson. Doubly-robust dynamic treatment regimen estimation for binary outcomes. *arXiv preprint arXiv:2203.08269*, 2022.

Alistair Johnson et al. MIMIC-III, A freely accessible critical care database. *Scientific Data*, 3(1), 2016.

Michael Kosorok and Eric Laber. Precision medicine. *Annual Review of Statistics and its Application*, 6:263–286, 2019.

Milan Kuzmanovic, Tobias Hatt, and Stefan Feuerriegel. Deconfounding temporal autoencoder: Estimating treatment effects over time using noisy proxies. In *ML4H*, 2021.

Siqi Liu, Kay Choong See, Kee Yuan Ngiam, Leo Anthony Celi, Xingzhi Sun, Mengling Feng, et al. Reinforcement learning for clinical decision support in critical care: comprehensive review. *Journal of Medical Internet Research*, 22(7):e18477, 2020.

Ying Liu et al. Deep reinforcement learning for dynamic treatment regimes on medical registry data. In *ICHI*, 2017.

Daniel J Luckett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, and Michael R Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 2019.

Erica Moodie, Thomas Richardson, and David Stephens. Demystifying optimal dynamic treatment regimes. *Biometrics*, 63(2):447–455, 2007.

Erica Moodie, Bibhas Chakraborty, and Michael S Kramer. Q-learning for estimating optimal dynamic treatment rules from observational data. *Canadian Journal of Statistics*, 40(4):629–645, 2012.

Susan Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

Susan Murphy, Mark van der Laan, James Robins, and CPPRG. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *Journal of the American Statistical Association*, 116(533):392–409, 2021.

Yilmazcan Oezyurt, Mathias Kraus, Tobias Hatt, and Stefan Feuerriegel. AttDMM: An attentive deep Markov model for risk scoring in intensive care units. *KDD*, 2021.

Niranjani Prasad, Li Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *Conference on Uncertainty in Artificial Intelligence*, 2017.

Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

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

James Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, 2004.

James Robins and Miguel Hernán. Estimation of the causal effects of time-varying exposures. In Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs, editors, *Advances in Longitudinal Data Analysis*, pages 553–599. Boca Raton, FL: CRC press, 2009.

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

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

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Kelly Speth and Lu Wang. Restricted sub-tree learning to estimate an optimal dynamic treatment regime using observational data. *Statistics in Medicine*, 40(26):5796–5812, 2021.

Kelly A. Speth, Alfred P. Yoon, Lu Wang, and Kevin C. Chung. Assessment of tree-based statistical learning to estimate optimal personalized treatment decision rules for traumatic finger amputations. *JAMA Network Open*, 3(2):e1921626–e1921626, 2020.

Yilun Sun and Lu Wang. Stochastic tree search for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 116(533):421–432, 2021.

Yebin Tao, Lu Wang, and Daniel Almirall. Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. *The Annals of Applied Statistics*, 12(3):1914–1938, 2018.

Anastasios Tsiatis et al. *Dynamic treatment regimes: Statistical methods for precision medicine*. Boca Raton, FL: CRC Press, 2019.

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.

Michael Wallace and Erica Moodie. Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics*, 71(3):636–644, 2015.

Michael Wallace, Erica Moodie, and David Stephens. Dynamic treatment regimen estimation via regression-based techniques: Introducing R package DTRreg. *Journal of Statistical Software*, 80(2):1–20, 2017. URL https://www.jstatsoft.org/index.php/jss/article/view/v080i02.

Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

Shirly Wang et al. MIMIC-extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III. In *CHIL*, 2020.

Gregory Yauney and Pratik Shah. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection. In *Machine Learning for Healthcare Conference*, pages 161–226. PMLR, 2018.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Yichi Zhang et al. Interpretable dynamic treatment regimes. *Journal of the American Statistical Association*, 113(524):1541–1549, 2018.

Ying-Qi Zhao et al. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598, 2015.

Yingqi Zhao and Donglin Zeng. Recent development on statistical methods for personalized medicine discovery. *Frontiers of Medicine*, 7(1):102–110, 2013.

Appendix A. Details on Baselines for Learning Optimal DTRs

Here, we describe the baseline methods against which we benchmark the perofrmance of DTR-CT and DTR-CF. Each of the baseline methods is embedded in Algorithm 1 to estimate the HTE in place of causal trees or a causal forest. We first describe classical methods to learn optimal DTRs using methods originating from statistics, and then describe how we can estimate optimal DTRs using machine learning methods.

A.1. Classical methods to Learning Optimal DTRs

Classical methods to learn optimal DTRs are, e. g., Q-learning (Moodie et al., 2012), DWOLS (Wallace and Moodie, 2015), and G-estimation (Robins, 2004). These typically makes use of linear regression-based models. At every time step $t = 1, \dots, T$, the potential outcome is modeled as the conditional expectation of the observed outcome given treatment and covariate history up until step t and given optimal treatments thereafter (Chakraborty and Moodie, 2013), i. e.,

$$\mathbb{E}[Y \mid \mathbf{H}_t = \mathbf{h}_t, A_t = a_t] = \beta_t^\top \mathbf{h}_t + a_t(\psi_t^\top \mathbf{h}_t). \tag{11}$$

Here, $\beta_t^\top \mathbf{H}_t$ is the time t treatment-free model that predicts the observed outcome in the absence of treatment in time t and $A_t(\psi_t^\top \mathbf{H}_t)$ is a treatment-interaction model, also known as a blip model, with which history covariates determine the heterogeneity in the effect of treatment at time t on the observed outcome.

The additive separation into a treatment-free model and a treatment-interaction model implies that treatment decisions are completely based on the latter. If the parameters β_t and ψ_t are known, the HTE at time t is given by $\tau(\mathbf{h}_t) = \psi_t^\top \mathbf{h}_t$. By the plug-in principle, its estimate is in turn $\hat{\tau}(\mathbf{h}_t) = \hat{\psi}_t^\top \mathbf{h}_t$. This gives an estimated optimal treatment decision

$$\hat{\delta}_t^{\text{opt}}(\mathbf{h}_t) = \mathbb{1}\{\hat{\psi}_t^\top \mathbf{h}_t > 0\}. \tag{12}$$

Note that although the parameters are estimated, these methods assume that the true model is low-dimensional, linear in parameters ψ_t, and has known functional form and variables. It also assumes that the same set of patient covariates are important for outcome prediction as for treatment effect heterogeneity and, thus, individualized treatment recommendations. In practice, these sets may be different and unknown.

A.2. Q-Learning

In Q-learning, the expected potential outcome is modeled as the conditional expectation of the outcome given treatment and covariate history (Moodie et al., 2012). At every time step $t = T, \dots, 1$ in the backward recursion, this conditional expectation is separated into a treatment-free model and the treatment-interaction model according to Eq. (11).

The blip parameters ψ_t are estimated with ordinary least squares. Specifically, the closed-form solution is given by the least squares equation

$$\hat{\psi}_t(\mathbf{H}_t, a_t) = \Big((a_t\mathbf{H}_t)^\top (a_t\mathbf{H}_t)\Big)^{-1} (a_t\mathbf{H}_t)^\top Y. \tag{13}$$

Given parameter estimates $\hat{\psi}_t = \hat{\psi}_t(\mathbf{H}_t, a_t)$ and using the plug-in principle, Q-Learning estimates the HTE at time step t as $\hat{\tau}(\mathbf{h}_t) = \hat{\psi}_t^\top \mathbf{h}_t$, which is a linear function in the parameters $\hat{\psi}_t$ and patient history $\mathbf{h}_t$. For each patient at every time step, the estimated optimal treatment decision is then given by Eq. (12).

A.3. Dynamic Weighted Ordinary Least Squares

Dynamic weighted ordinary least squares (DWOLS) extends Q-learning by weighting observations with the inverse propensity score estimate $\hat{\pi}_t^{-1} = (\hat{\pi}(\mathbf{h}_t))^{-1}$ (Wallace and Moodie, 2015; Wallace et al., 2017). To estimate the blip parameter for time step $t = 1, \ldots, T$, the observed outcome Y is regressed on $(\mathbf{H}_t, a_t\mathbf{H}_t)$. The closed-form least squares solution is

$$\hat{\psi}_t(\mathbf{H}_t, a_t) = \Big((a_t\mathbf{H}_t)^\top(\hat{\pi}_t^{-1}a_t\mathbf{H}_t)\Big)^{-1}(a_t\mathbf{H}_t)^\top(\hat{\pi}_t^{-1}Y). \tag{14}$$

The HTE is then estimated as $\hat{\tau}(\mathbf{h}_t) = \hat{\psi}_t^\top \mathbf{h}_t$, which, like in Q-learning, is a linear function in parameters and patient history. Also as for Q-learning, we use the same sets of history covariates to estimate the two models involved in DWOLS. The only difference to Q-learning is that by weighing observations with the inverse propensity score, DWOLS is doubly robust against model mis-specification.

A.4. G-Estimation

Similar to Q-learning and DWOLS, G-estimation is a semi-parametric regression-based method to sequentially estimate HTEs using backward recursion (Robins, 2004). However, instead of weighting observations with the inverse propensity score, like DWOLS, it directly incorporates the estimated propensity score $\hat{\pi}(\mathbf{h}_t)$ in the estimating equation, given by

$$\hat{\psi}_t(\mathbf{H}_t, a_t) = \Big(((a_t - \hat{\pi}_t)\mathbf{H}_t)^\top(a_t\mathbf{H}_t)\Big)^{-1}((a_t - \hat{\pi}_t)\mathbf{H}_t)^\top Y. \tag{15}$$

Using blip parameters obtained from the above estimating equation, the HTE estimate is given by $\hat{\tau}(\mathbf{h}_t) = \hat{\psi}_t^\top \mathbf{h}_t$.

A.5. CART

CART (Breiman et al., 1984) (or simply, decision trees) use a splitting criterion that aggregates homogeneous patients in terms of their outcomes. We fit two separate regression trees, CART1 and CART0, for the conditional expectation of the observed outcome among the treated and the control groups, given the history and treatment, at every time step t. The estimated HTE is then given by their difference, i. e.,

$$\hat{\tau}_t^{(\text{CART})}(\mathbf{h}_t) = \hat{\mathbb{E}}_{\text{CART1}}\left[Y \mid \mathbf{H}_t = \mathbf{h}_t, A_t = 1\right] - \hat{\mathbb{E}}_{\text{CART0}}\left[Y \mid \mathbf{H}_t = \mathbf{h}_t, A_t = 0\right]. \tag{16}$$

A.6. k-NN Regression

With k-NN regression, we estimate the HTE at the time step t as the difference in conditional expectation of the observed outcomes for the treated and control groups, given the patient histories:

$$\hat{\tau}_t^{(k\text{-NN})}(\mathbf{h}_t) = \hat{\mathbb{E}}_{k\text{-NN1}}\left[Y \mid \mathbf{H}_t = \mathbf{h}_t, A_t = 1\right] - \hat{\mathbb{E}}_{k\text{-NN0}}\left[Y \mid \mathbf{H}_t = \mathbf{h}_t, A_t = 0\right]. \tag{17}$$

Here, k-NN1 and k-NN0 denote k-nearest neighbor (NN) regression models estimated on the treatment and control groups, respectively. The models are fitted on the k-nearest patients in each group within the neighbourhood of the patient of interest in terms of history covariate values $\mathbf{h}_t$.

Appendix B. Details on Data Generating Processes

B.1. Scenario 1: Linear Interactions

The true treatment-free functions for $t = 1, 2, 3$ are constructed via

$$\phi_1(\mathbf{X}_1) \coloneqq X_{11} + X_{12} + X_{13} + X_{14}, \tag{18}$$
$$\phi_2(\mathbf{X}_2) \coloneqq X_{21} + X_{22} - X_{23} + X_{24}, \tag{19}$$
$$\phi_3(\mathbf{X}_3) \coloneqq X_{31} + X_{32} - X_{33} + X_{34}. \tag{20}$$

B.2. Scenario 2: Non-Linear Interactions

The true treatment-free functions for $t = 1, 2, 3$ are constructed non-linearly via

$$\phi_t(\mathbf{X}_1) \coloneqq 5 + X_{19}^2 - 3\sin(C_2^2 + X_{11} - X_{12} + X_{13}), \tag{21}$$
$$\phi_t(\mathbf{X}_2) \coloneqq \mu_1(\mathbf{X}_1) + X_{29}^2 - 3\sin(X_{21} - X_{22} + X_{23}), \tag{22}$$
$$\phi_t(\mathbf{X}_3) \coloneqq \mu_2(\mathbf{X}_2) + X_{39}^2 - 3\sin(X_{31} - X_{32} + X_{33}). \tag{23}$$

See the main paper text and codes at our repository (https://github.com/tbluemlein/DTR) for details on how the data and HTEs are constructed.

B.3. Details on Implementation of Machine Learning Baselines

CART. We set the minimum number of observations required in each leaf to 10. This is half of the required number used in DTR-CT because we train two separate trees, whereas in DTR-CT, we train a single tree on treated and control patients at each time t.

k-NN Regression. We set $k = 10$ in each estimator, which corresponds to the minimum number of observations required in each causal tree in DTR-CT, i. e., 20.

DTR-CT. We use 50 % of the patient population as training set and 50 % as estimation set at each step t. The minimum number of treated and control patients in each leaf of every causal tree is set to 20 in total, consisting of 10 patients from each subgroup. To prevent overfitting, the maximum number of buckets used for splitting is set to 20. We run 5-fold cross-validation on Eq. (4) to optimize the pruning of the decision trees.

DTR-CF. We estimate each causal forest per time step in DTR-CF as an ensemble of 500 causal trees. We set the minimum number of treated and control patients in each leaf of every causal tree to 20 in total, consisting of 10 patients from each subgroup.

For details, see the codes at our repository (https://github.com/tbluemlein/DTR).

B.4. Causal sensitivity analysis

To test the validity of Assumption 4 for our empirical application to real-world data, we perform a causal sensitivity analysis. We follow current state-of-the-art and use the method

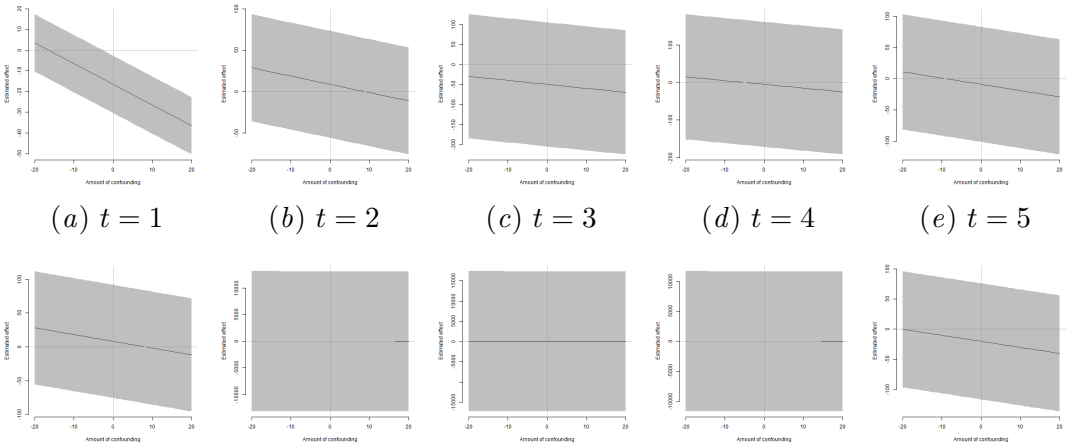
proposed by (Blackwell, 2014). The method allows us to evaluate the effect of potential unmeasured confounders on the estimated treatment effects while varying both the strength and the direction of the confounding.

We run the causal sensitivity analysis using the R package `causalsens`.[4]. The package requires specifying a propensity score model and an outcome model. For each time step, we specify the propensity score model as the logistic regression model used in our empirical application and fit it to the observational data up until that time step. For the outcome model, the package currently only supports linear models estimated with ordinary least squares, possibly with weighted outcomes. We thereby specify the outcome model per time step as Eq. (11) and fit it to the observational data. Here, we use inverse propensity score weighting of the outcome with weights predicted from the fitted logistic regression model. We note that although we cannot perform the causal sensitivity analysis for our proposed approaches DTR-CT and DTR-CF, their estimated treatment effects may be less susceptible to unmeasured confounding than the estimated treatment effects reported in this causal sensitivity analysis. This is the case the causal tree and causal forest regression models utilized in DTR-CF and DTR-CT detect which covariates to control for and their possible non-linear relationships with the outcome.

We follow Blackwell (2014) and demonstrate the impact of potential unmeasured confounding by plotting the estimated treatment effects as a function of the amount of confounding and as a function of the amount of the unexplained variance in the potential outcomes. We estimate the average treatment effects on the treated rather than the HTEs as the former would require one plot per heterogenous individual in the data. Our causal sensitivity analysis thereby measures the average sensitivity to unmeasured confounding among the patient-level HTEs per time step.

The results suggest that most of the average treatment effect estimates across the time steps do not vary with potential unmeasured confounding (Fig 5). In particular, only the estimate for the first time step seem somewhat sensitive to unmeasured confounding. The same result appears to hold for the estimated average treatment effects as a function of the unexplained variance in the potential outcomes (Fig. 6). Overall, the causal sensitivity analysis suggest that our empirical application do not suffer from worrying amounts of unmeasured confounding.

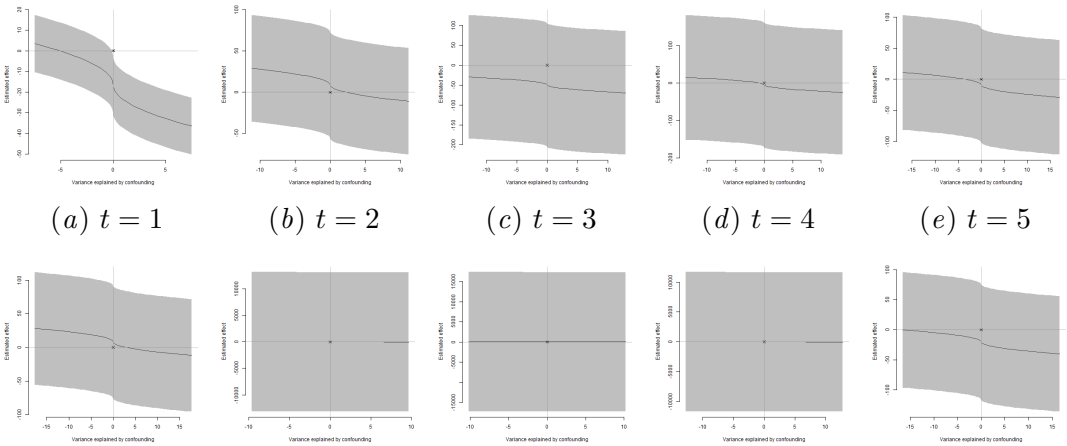
4. Details are available on the official web page: https://CRAN.R-project.org/package=causalsens



$(a)\ t = 1$ $(b)\ t = 2$ $(c)\ t = 3$ $(d)\ t = 4$ $(e)\ t = 5$

$(f)\ t = 6$ $(g)\ t = 7$ $(h)\ t = 8$ $(i)\ t = 9$ $(j)\ t = 10$

Figure 5: Causal sensitivity analysis for unmeasured confounding with inverse propensity score weighting of the outcome. The plots show the estimated average treatment effect on the treated per time step as a function of the amount of confounding in the potential outcomes.



$(a)\ t = 1$ $(b)\ t = 2$ $(c)\ t = 3$ $(d)\ t = 4$ $(e)\ t = 5$

$(f)\ t = 6$ $(g)\ t = 7$ $(h)\ t = 8$ $(i)\ t = 9$ $(j)\ t = 10$

Figure 6: Causal sensitivity analysis for unmeasured confounding with inverse propensity score weighting of the outcome. The plots show the estimated average treatment effect on the treated per time step as a function of the amount of the unexplained variance in the potential outcomes.