Received <day> <Month>, <year>; | Revised <day> <Month>, <year>; | Accepted <day> <Month>, <year>

DOI: xxx/xxxx

ARTICLE TYPE

Restricted Sub-Tree Learning (ReST-L) to Estimate an Optimal Dynamic Treatment Regime Using Observational Data

Kelly Speth | Lu Wang

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, United States

Correspondence
Lu Wang, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, 48109.
Email: luwang@umich.edu

Abstract

Dynamic treatment regimes (DTRs), consisting of a sequence of tailored treatment decision rules that span multiple stages of care, present a unique opportunity in our drive toward personalized medicine. Given that estimation of optimal DTRs is often exploratory and communication with clinicians is vital, robust and flexible methods that yield interpretable results are needed. Tree-based methods utilizing a purity measure defined on the full set of covariates have enjoyed much success in meeting this goal. Often, however, it is necessary for clinical, practical or ethical reasons to restrict certain covariates that should be used when making treatment decisions. Herein we present Restricted Sub-Tree Learning (ReST-L), a flexible and robust, sub-tree-based method to estimate an optimal multi-stage multi-treatment DTR that enables restrictions to the set of prespecified candidate tailoring variables. ReST-L employs a purity measure derived from an augmented inverse probability weighted estimator for the counterfactual mean outcome, using observational data to build multi-stage decision trees that are restricted in sub-tree spaces defined by the corresponding prescriptive covariates. We show that ReST-L is able to correctly estimate the optimal DTR searching over a large number of variables with relatively small sample sizes and improves upon competing estimation methods. We demonstrate the utility of ReST-L to estimate a two-stage fluid resuscitation strategy for patients admitted to an intensive care unit with acute emergent sepsis.

KEYWORDS:
Adaptive interventions; Personalized medicine; Restricted optimization; Tailoring variables; Tree-based statistical learning

1 | INTRODUCTION

There is a drive in the healthcare field toward evidence-based and personalized medicine, which has the potential to both improve patient outcomes, lower costs, and allocate healthcare resources in more efficient ways. One such avenue to achieve this goal is through dynamic treatment regimes, which have become of great recent interest in several medical specialties including oncology, as well as in social and clinical psychology and behavioral health. Dynamic treatment regimes,[1,2,3] also known commonly as DTRs, individualized treatment rules, or adaptive interventions, represent multi-stage, prescribed treatment sequences that

[0]**Abbreviations:** DTR, dynamic treatment regime; MIMIC, Medical Information Mart

are tailored to the individual based on their baseline and time-varying characteristics and are a vehicle to operationalize the manner in which patient care for chronic diseases is delivered in practice more efficiently.

A primary statistical objective in the field of DTRs is to estimate stage-specific decision rules that will optimize the expected long-term counterfactual outcomes of patients when applied across the population of interest. Given the potential value of DTRs for both improving long-term patient outcomes and optimizing the allocation of resources needed for patient care, there are understandably a myriad of methods that have been developed to estimate optimal DTRs. Existing methods are vast and may be classified by their dependence on parametric or semi-parametric, or nonparametric assumptions. Tree-based methods that offer flexibility and robustness in estimation are desirable given, often, an abundance of observational data, as well as a high degree of uncertainty with regard to the complex relationships among variables. Additionally, because optimal DTR estimation is exploratory in nature and communication with clinicians is crucial, methods with interpretable results are particularly favored. Tree-based methods informing optimal DTR construction in a single stage and/or with binary treatment decisions[4,5,6,7,8] have been expanded to accommodate a multi-stage and multi-treatment setting. Zhang et al.[9] propose a robust and flexible method yielding interpretable estimated regimes in the form of a decision list, but these lists incorporate only two covariates per rule and grow unidimensionally. Tao et al.[10] introduce tree-based reinforcement learning (T-RL), a semi-parametric approach combining the flexibility of a decision tree (e.g., Breiman et al.[11]) with a purity measure that is derived from a doubly-robust augmented inverse probability weighted estimator of the counterfactual mean outcome.[8] Although T-RL has desirable properties and is able to accommodate multiple treatments across multiple stages, the algorithm requires all observed covariates to be considered as possible tailoring variables in an optimal DTR, which is unlikely to exclusively occur in practice.

Consider the case of a patient admitted to the intensive care unit (ICU) with acute emergent sepsis, a clinical syndrome that is associated with one of the highest rates of mortality among conditions commonly treated in the ED.[12] Current practice recommends treating all patients with early liberal fluid resuscitation.[13] However, this recommendation is given with a stated "low quality of evidence" due to the fact that results across studies have been inconsistent with indirect evidence, imprecise results, and a likelihood of bias. When observational data are used to elucidate a causal relationship, all measured covariates may be used to evaluate the treatment assignment mechanism. Many of these covariates, including ethnic/racial identity or type of insurance, for example, would, for a variety of reasons, be considered inappropriate to include as a potential tailoring variable to make a treatment decision. We therefore seek a flexible and robust causal method with desirable statistical properties and interpretable results to estimate an optimal multi-stage DTR that is restricted in its prescriptive covariates.

In this manuscript we propose Restricted Sub-Tree Learning (ReST-L), which utilizes a decision tree framework but restricts the covariate space, according to subject-matter knowledge, to build an estimated, optimal DTR based on a sub-tree, a topic that to our knowledge has not yet been explored in the statistical literature. In order to determine the binary splits of the covariate space that define the decision tree at each stage, we propose a purity measure derived using the AIPW estimator of the counterfactual mean outcome, but with an important modification. Whereas the AIPW estimator of the counterfactual mean outcome is estimated using the full set of observed covariates, which is necessary for a causal interpretation, we restrict the set of candidate tailoring variables to only those deemed reasonable based on clinical or scientific expertise. In simulation studies we demonstrate that, when clinical knowledge substantiates consideration of only a subset of covariates as candidate tailoring variables but other covariates may define the treatment assignment mechanism or may be related to the outcome, ReST-L provides a flexible, semi-parametric analysis approach with interpretable estimates of an optimal, multi-stage DTR that demonstrates superior performance to existing methods.

2 | MATHEMATICAL FORMULATION AND ASSUMPTIONS

2.1 | Notation and Formulation

Suppose one of K_j treatments ($k_j = 1, ..., K_j$; $K_j \geq 2$) is administered to every subject $i = 1, ..., n$ at each of $j = 1, ..., J$ stages. The actual treatment received by the i-th patient at stage j is denoted $A_{j,i}$. As is customary, we use the convention of using a capital letter to refer to the random variable and lowercase to refer to a realized value. For simplicity, we omit the subscript i from future notation when no confusion exists. Let us denote the j-th stage covariates that are observed and available when making the j-th treatment decision as $\mathbf{X}_j$. Assuming that only a subset of covariates among $\mathbf{X}_j$ may be used to define the j-th stage decision rule, we distinguish between $\mathbf{X}_{\text{sub,j}}$ and $\mathbf{X}_{\text{sub,j}}{}^{C} = \mathbf{X}_j \setminus \mathbf{X}_{\text{sub,j}}$, where $\mathbf{X}_{\text{sub,j}}$ represents a p_j-dimensional vector of multi-scale data ($p_j \geq 1$) corresponding to measured covariates that a clinician would consider as candidates to include in a treatment decision rule at stage j. Conversely, $\mathbf{X}_{\text{sub,j}}{}^{C}$ is a q_j-dimensional vector of multi-scale data ($q_j \geq 1$)

corresponding to the set of measured covariates that may *not* be included in a clinical treatment decision rule. In the context of our application example to determine an optimal treatment regime to treat patients with acute emergent sepsis, $\mathbf{X}_{\text{sub}}$ includes variables such as the Elixhauser comorbidity index and prior treatment with vasopressors or mechanical ventilation–all of which a clinician would consider as possible tailoring variables–whereas $\mathbf{X}^C_{\text{sub}}$ includes the patient's racial/ethnic identity, gender, the time of year in which care was delivered, and other variables that a clinician would not use to assign treatment. After each j-th treatment stage, measurements are made on a set of covariates $\mathbf{X}_{j+1}$, which may also include an intermediate outcome, Y_j. Such an intermediate outcome Y_j may reflect a certain response from previous treatments or may be a function of the previous treatment history and other observed covariates, and may be used to determine treatment for the $(j+1)$-th stage. Thus, Y_j may also be an element of $\mathbf{X}_{\text{sub},(j+1)}$. Following convention, we use overbar to denote history, i.e., all observations collected at stages on or before the jth stage. For example, $\overline{\mathbf{X}}_j = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_j)$ represents all covariate data collected and available to the clinician prior to the jth treatment decision. Similarly, $\overline{A}_{j-1}$ represents the set of all treatments received prior to the jth treatment decision, i.e., $\overline{A}_{j-1} = (A_1, A_2, ..., A_{j-1})$, and $\overline{Y}_{j-1} = (Y_1, Y_2, ..., Y_{j-1})$ represents the set of all intermediate outcomes observed prior to the jth treatment decision. Suppose the final outcome of interest is $Y = h(Y_1, Y_2, ..., Y_J)$, which may be a function of stage-specific intermediate outcomes $Y_1, Y_2, ..., Y_J$, is assumed to be continuous and approximiately normally distributed. Here $h(\cdot)$ represents some clinically-relevant, prespecified function (e.g., sum or last value). The full history prior to the decision at stage j is then expressed as $\mathbf{H}_j = (\overline{A}_{j-1}, \overline{\mathbf{X}}_j)$. $\mathbf{H}_{\text{sub,j}} = (\overline{A}_{j-1}, \overline{\mathbf{X}}_{\text{sub,j}})$ includes the full treatment history prior to the treatment decision at stage j and covariate history only from the subset $\overline{\mathbf{X}}_{\text{sub,j}}$ that may be used in a clinical decision rule. Using a similar convention, $\mathbf{H}^C_{\text{sub,j}} = \overline{\mathbf{X}}^C_{\text{sub,j}}$ includes covariate history through stage j for variables not considered for a treatment regime. Next let $\mathbf{g} = (g_1, g_2, ..., g_J)$ denote a J-stage DTR. Each stage-specific decision rule g_j is a function only of covariates that can be used to make treatment decisions at each stage, i.e., $g_j : \mathbf{H}_{\text{sub, j}} \rightarrow A_j$.

As we are interested in making a causal claim related to an estimated optimal DTR, we employ Rubin's potential outcome framework.[14] At stage J we let $Y^*(A_1, ..., A_{J-1}, a_J)$, or simply $Y^*(a_J)$, denote the counterfactual outcome, also known as a potential outcome, for a patient treated with $a_J \in \mathcal{A}_J$ conditional on prior treatment history $\overline{A}_{J-1}$. Notably, only one counterfactual outcome–the one consistent with the treatment actually received–will be observed. In the context of our estimation problem we can similarly define $Y^*\{\mathbf{g}(\mathbf{H}_{\text{sub}})\}$ as the counterfactual outcome under the multi-stage DTR $\mathbf{g}(\mathbf{H}_{\text{sub}})$. As mentioned above, only one counterfactual outcome will be observed, although in this case the observed counterfactual will be the potential outcome consistent with the DTR followed by the individual. We measure the performance of a multi-stage DTR, $\mathbf{g}(\mathbf{H}_{\text{sub}})$, using the counterfactual mean outcome $E[Y^*\{\mathbf{g}(\mathbf{H}_{\text{sub}})\}]$, the higher the better by convention, and define the optimal DTR $\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})$ as the one that satisfies

$$E[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] \geq E[Y^*\{\mathbf{g}(\mathbf{H}_{\text{sub}})\}]$$

for all $\mathbf{g}(\mathbf{H}_{\text{sub}}) = (g_1, g_2, ..., g_J)^T \in \mathcal{G}_{\text{sub}}$, where $\mathcal{G}_{\text{sub}}$ is the class of all potential regimes constructed using $\mathbf{H}_{\text{sub}}$ only. Our statistical goal, therefore, can be summarized as follows: to estimate an interpretable, optimal, J-stage treatment regime, $\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})$, from observational data such that, if all patients were to be assigned to multi-stage treatment using this regime, the expected counterfactual outcome of our population of interest would be maximized: $\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}}) = \text{argmax}_{\mathbf{g}\in\mathcal{G}_{\text{sub}}} E[Y^*\{\mathbf{g}(\mathbf{H}_{\text{sub}})\}]$.

2.2 | Link to Observed Data

The above optimization objective features $Y^*\{\mathbf{g}(\mathbf{H}_{\text{sub}})\}$, the counterfactual outcome under DTR $g(\mathbf{H}_{\text{sub}})$; however, as is widely known in causal inference, only one of the potential outcomes is observed, making estimation of a causal effect impossible without a series of assumptions. To proceed with estimation of an optimal DTR under Rubin's potential outcomes framework, we make three foundational assumptions: consistency, positivity, and no unmeasured confounders (NUCA). Under an assumption of consistency, the potential outcome under the observed treatment agrees with that of the observed outcome Y. For a single stage treatment, we can express this as: $Y = \sum_{a=1}^{K} Y^*(a)\mathcal{I}(A = a)$, where $\mathcal{I}(\cdot)$ is an indicator function that returns a value of 1 if the argument is true and a value of 0 otherwise. Consistency further assumes that there is no interference between units, which means that one patient's observed and counterfactual outcomes are independent of the treatment(s) of all other patients. Following positivity, $0 < \tau < Pr(A_i = a|\mathbf{H}_i) < 1$ for all $a \in \mathcal{A}, \mathbf{H}_i \in \boldsymbol{\mathcal{H}}$, where τ is a positive constant, i.e., the probability to be assigned to each possible treatment is bounded away from 0. Finally, we assume that there are no unmeasured confounders, i.e., data on all variables associated with both the assignment of treatment A and the outcome Y have been observed. That is, given history $\mathbf{H}$, $Y^*_1(1), ..., Y^*_K(K) \perp\!\!\!\perp A|\mathbf{H}$, where $\perp\!\!\!\perp$ denotes statistical independence.

In contrast to previous work with tree-based reinforcement learning,[15,10] our interest is to estimate a $\mathbf{g}^{\text{opt}}$ that is based only on covariates in $\mathbf{H}_{\text{sub}}$. Because $Y^*\{g(\mathbf{H}_{\text{sub}})\} = \sum_{a=1}^{K} Y^*(a)\mathcal{I}\{g(\mathbf{H}_{\text{sub}}) = a\}$, the optimal decision rule g for a single stage can be expressed as $g^{\text{opt}}(\mathbf{H}_{\text{sub}}) = \text{argmax}_{g\in\mathcal{G}_{\text{sub}}} E\left[\sum_{a=1}^{K} Y^*(a)\mathcal{I}\{g(\mathbf{H}_{\text{sub}}) = a\}\right]$. After taking an iterated expectation and conditioning on history $\mathbf{H}$, and in accordance with our assumptions of consistency, positivity, and NUCA, we can express the optimal decision rule as follows:

$$g^{\text{opt}}(\mathbf{H}_{\text{sub}}) = \text{argmax}_{g\in\mathcal{G}_{\text{sub}}} E[Y^*\{g(\mathbf{H}_{\text{sub}})\}] = \text{argmax}_{g\in\mathcal{G}_{\text{sub}}} E_{\mathbf{H}}\left[\sum_{a=1}^{K} E\{Y|A = g(\mathbf{H}_{\text{sub}}) = a, \mathbf{H}\}\right]$$

3 | RESTRICTED SUB-TREE LEARNING (REST-L)

3.1 | Estimator of Counterfactual Mean Outcome under Regime $g(\mathbf{H}_{\text{sub}})$ for a Single Treatment Stage

Consider estimation of the optimal decision rule $g^{\text{opt}}(\mathbf{H}_{\text{sub}})$ for a single stage with K possible treatments: $g^{\text{opt}} : \boldsymbol{\mathcal{H}}_{\text{sub}} \rightarrow \{1, 2, ..., K\}$. Define C as a compatibility indicator, with $C = \sum_{a=1}^{K} \mathcal{I}(A = a) \cdot \mathcal{I}\{g(\mathbf{H}_{\text{sub}}) = a\}$, which is equivalent to $\mathcal{I}\{A = g(\mathbf{H}_{\text{sub}})\}$, meaning that the actual treatment received is consistent with the treatment assigned by rule $g(\mathbf{H}_{\text{sub}})$. Next define $\pi_a(\mathbf{H}) = Pr(A = a|\mathbf{H})$ as the propensity score for treatment assignment, noticing that this potentially depends on all variables in $\mathbf{H}$–not just variables in $\mathbf{H}_{\text{sub}}$. Also, denote $\pi_C(\mathbf{H})$ as the probability of receiving treatment consistent with $g(\mathbf{H}_{\text{sub}})$. Assuming we have observational data, we would posit a propensity model $\pi_a(\mathbf{H}; \gamma)$, e.g., using multinomial logistic regression, to estimate γ. We see that:

$$\pi_C(\mathbf{H}) = Pr(C = 1|\mathbf{H}) = E(C|\mathbf{H}) = E\Big[\sum_{a=1}^{K} \mathcal{I}(A = a) \cdot \mathcal{I}\{g(\mathbf{H}_{\text{sub}}) = a\}|\mathbf{H}\Big] = \sum_{a=1}^{K} \pi_a(\mathbf{H}) \cdot \mathcal{I}\{A = g(\mathbf{H}_{\text{sub}}) = a\}$$

Under the three assumptions in Section 2.2, consider the IPW estimator of $E[Y^*\{g(\mathbf{H}_{\text{sub}})\}]$, i.e., $\hat{E}[Y^*\{g(\mathbf{H}_{\text{sub}})\}] = \mathbb{P}_n\left\{\frac{C \cdot Y}{\hat{\pi}_C(\mathbf{H}_i; \hat{\gamma})}\right\}$, where C and $\pi_C(\mathbf{H})$ are defined above, $\mathbb{P}_n(\cdot)$ represents the empirical mean operator evaluated over all patients i, and Y represents our outcome of interest. Under an assumption of consistency and positivity:

$$E\left[\frac{C \cdot Y}{\hat{\pi}_C(\mathbf{H}; \hat{\gamma})}\right] = E\left[\frac{C}{\hat{\pi}_C(\mathbf{H}; \hat{\gamma})} Y^*\{g(\mathbf{H}_{\text{sub}})\}\right]$$

Taking an iterated expectation conditional on $\mathbf{H}$ and under the assumption of NUCA, the above is equivalent to:

$$E_{\mathbf{H}}\left[E\Big[\frac{C}{\hat{Pr}(C = 1|\mathbf{H})} Y^*\{g(\mathbf{H}_{\text{sub}})\}|\mathbf{H}\Big]\right] = E_{\mathbf{H}}\left[E\left[\frac{\mathcal{I}\{A = g(\mathbf{H}_{\text{sub}})\}}{\hat{Pr}\{A = g(\mathbf{H}_{\text{sub}})|\mathbf{H}\}}|\mathbf{H}\right] E[Y^*\{g(\mathbf{H}_{\text{sub}})\}|\mathbf{H}]\right]$$

If $\hat{\pi}_C(\mathbf{H})$ is correctly specified, $E\left[\frac{\mathcal{I}\{A = g(\mathbf{H}_{\text{sub}})\}}{\hat{Pr}\{A = g(\mathbf{H}_{\text{sub}})|\mathbf{H}\}}|\mathbf{H}\right] = 1$, which demonstrates that an IPW-style estimator is consistent in large samples for estimating the counterfactual mean outcome $E[Y^*\{g(\mathbf{H}_{\text{sub}})\}]$ under a regime $g(\mathbf{H}_{\text{sub}})$:

$$\hat{E}[Y^*\{g(\mathbf{H}_{\text{sub}})\}] = \mathbb{P}_n\left[\frac{C \cdot Y}{\hat{\pi}_C(\mathbf{H}; \hat{\gamma})}\right] \rightarrow^p E[Y^*\{g(\mathbf{H}_{\text{sub}})\}]$$

However, because the IPW estimator of the counterfactual mean can quickly become unstable as the number of treatment stages increases, estimation can be improved by utilizing a doubly robust estimator of the counterfactual mean, also known as the augmented inverse probability weighted (AIPW) estimator. It has been shown that $\mathbb{P}_n\{\hat{\mu}_a^{\text{AIPW}}(\mathbf{H})\}$ is a consistent estimator for $E\{Y^*(a)\}$ (Tao et al.,[10] and others), where $\hat{\mu}_a^{\text{AIPW}}(\mathbf{H})$ is defined as follows, with $\hat{\mu}_a(\mathbf{H}) = E(Y|A = a, \mathbf{H})$ representing the conditional mean model:

$$\hat{\mu}_a^{\text{AIPW}}(\mathbf{H}) = \frac{\mathcal{I}(A = a)}{\hat{\pi}_a(\mathbf{H})} Y + \left\{1 - \frac{\mathcal{I}(A = a)}{\hat{\pi}_a(\mathbf{H})}\right\} \hat{\mu}_a(\mathbf{H})$$

Now considering the decision rule $g(\mathbf{H}_{\text{sub}})$, we propose to extend the AIPW estimator to estimate the counterfactual mean under regime $g(\mathbf{H}_{\text{sub}})$, i.e., $\hat{E}[Y^*\{g(\mathbf{H}_{\text{sub}})\}]$, as $\mathbb{P}_n\left[\sum_{a=1}^{K} \hat{\mu}_a^{\text{AIPW}}(\mathbf{H})\, \mathcal{I}\{g(\mathbf{H}_{\text{sub}}) = a\}\right]$, which can also be expressed as:

$$\hat{E}[Y^*\{g(\mathbf{H}_{\text{sub}})\}] = \mathbb{P}_n\left[\frac{C}{\hat{\pi}_C(\mathbf{H})} Y + \left\{1 - \frac{C}{\hat{\pi}_C(\mathbf{H})}\right\} \hat{\mu}_C(\mathbf{H})\right]$$

where both C and $\pi_C(\mathbf{H})$ are as defined above, and $\hat{\mu}_C(\mathbf{H}) = \hat{E}\{Y|A = g(\mathbf{H}_{\text{sub}}) = a, \mathbf{H}\}$. The AIPW estimator of the counterfactual mean outcome for treatment $A = a$ under regime $g(\mathbf{H}_{\text{sub}})$ is then expressed as:

$$\mathbb{P}_n\left[\frac{\mathcal{I}\{A = g(\mathbf{H}_{\text{sub}}) = a\}}{\hat{\pi}_C(\mathbf{H})}Y + \left\{1 - \frac{\mathcal{I}\{A = g(\mathbf{H}_{\text{sub}}) = a\}}{\hat{\pi}_C(\mathbf{H})}\right\}\hat{\mu}_C(\mathbf{H})\right]$$

It is important here to highlight the fact that the full set of covariates, **H**, are used in both the propensity and conditional mean models. This AIPW estimator is doubly robust in the sense that it will provide a consistent estimator of the counterfactual mean outcome under regime $g(\mathbf{H}_{\text{sub}})$ if either the models of $\hat{\pi}_C(\mathbf{H})$ or the conditional mean outcome model $\hat{\mu}_C(\mathbf{H})$ is correctly specified.

3.2 | Tree-based Estimation

With the above knowledge, we propose ReST-L, a new statistical learning procedure akin to the classification and regression tree[11] (CART), to estimate $g(\mathbf{H}_{\text{sub}})$. One important feature of CART methods, which will also be an important component of ReST-L, is the purity measure. A purity measure is used to quantify the degree of similarity–or "purity"–of observations with respect to a target variable. Specifically, the measured purity is used to determine binary covariate splits, mimicking the branching of a tree, such that observations within each "leaf" node are relatively homogeneous with respect to a target variable–and then to use the estimated partition of the covariate space to predict the target variable for a set of new observations. The process of splitting nodes of the tree into binary partitions of the covariate space continues until the pre-specified depth of the tree is achieved or until the improvement in the purity falls below a pre-specified level. Examples of purity measures frequently used with CART include entropy, the Gini index, and the sum of squared prediction errors.[16] A similarity between CART and ReST-L, as introduced above, includes the fact that a partition of the covariate space is made such that observations within each subset are relatively homogeneous. A notable difference, however, is the fact that the target of estimation for ReST-L is an optimal decision rule, which determines optimal treatment assignment based on observed covariates, but is not directly observed. Furthermore, ReST-L, in contrast to CART methods, rests within the causal inference framework. Therefore, we propose for ReST-L a new purity measure suitable for our goal of estimating a decision rule based on only a subset of covariates while also preserving a causal interpretation. Specifically, we exploit the consistent, large-sample, doubly-robust AIPW estimator of the counterfactual mean outcome for a decision rule based only on a subset of covariates introduced in Section 3.1. We define our purity measure, $\mathcal{P}(\Omega, \omega)$, represented by the binary partition created by split ω of node Ω as follows:

$$\mathcal{P}(\Omega, \omega) = \max_{a_1, a_2 \in \mathcal{A}} \; \mathbb{P}_n\left[\sum_{a=1}^{K} \hat{\mu}_a^{\text{AIPW}}(\mathbf{H}) \;\; \mathcal{I}\{A = g_{\omega, a_1, a_2}(\mathbf{H}_{\text{sub}}) = a\} \;\; \mathcal{I}(\mathbf{H}_{\text{sub}} \in \Omega)\right],$$

where g_{ω,a_1,a_2} denotes a decision rule such that patients in ω are assigned treatment a_1 while patients in the complementary set ω^C are assigned to a_2, with $a_1 \neq a_2$. Note that the purity measure is constructed such that all covariates **H** may be used when constructing the AIPW estimator of the counterfactual mean outcome under regime $g(\mathbf{H}_{\text{sub}})$, but only a subset, $\mathbf{H}_{\text{sub}}$, is selected as potential tailoring variables. Using this purity measure, ReST-L is implemented as described in Section 3.4.

3.3 | Estimation for Multiple Treatment Stages

We now extend ReST-L to a setting with multiple treatment stages, $j = 1, ..., J$, with 2 or more treatment options per stage, i.e., $K_j \geq 2$. Because of the potential for confounding by indication, a problem that can introduce substantial bias in a multi-stage estimation, ReST-L is implemented recursively using backward induction,[17] beginning with estimation of the final stage. Backward induction is an iterative process of reasoning backward in time, first determining the optimal solution at the final stage, followed by successively earlier stages, and has become the standard for determining the optimal solution in a multi-stage problem in a causal framework.

We first consider estimation of the decision rule for the final, J-th, stage. We perform this estimation in the same manner in which we estimate a single stage decision rule. Following our exposition in section 3.1,

$$\hat{\mu}_{J,a_J}^{\text{AIPW}}(\mathbf{H}_J) = \frac{\mathcal{I}(A_J = a_J)}{\hat{\pi}_{J,a_J}(\mathbf{H}_J)}Y + \left\{1 - \frac{\mathcal{I}(A_J = a_J)}{\hat{\pi}_{J,a_J}(\mathbf{H}_J)}\right\}\hat{\mu}_{J,a_J}(\mathbf{H}_J)$$

where $\hat{\mu}_{J,a_J}(\mathbf{H}_J) = \hat{E}[Y|A_J = a_J, \mathbf{H}_J]$. The estimator of the J-th stage counterfactual mean outcome for $A_J = a_J$ under regime $g_J(\mathbf{H}_{\text{sub,J}})$ can then be expressed as:

$$\mathbb{P}_n\left(\frac{\mathcal{I}\{A_J = g_J(\mathbf{H}_{\text{sub,J}}) = a_J\}}{\hat{\pi}_{J,C_J}(\mathbf{H}_J)}Y + \left[1 - \frac{\mathcal{I}\{A_J = g_J(\mathbf{H}_{\text{sub,J}}) = a_J\}}{\hat{\pi}_{J,C_J}(\mathbf{H}_J)}\right]\hat{\mu}_{C_J}(\mathbf{H}_J)\right)$$

where $\hat{\mu}_{C_J}(\mathbf{H}_J) = E\{Y|A_J = g_J(\mathbf{H}_{\text{sub,J}}) = a_J, \mathbf{H}_J\}$. Likewise, we define the purity measure for the J-th stage decision rule $g_J(\mathbf{H}_{\text{sub}})$ under a binary split ω (and ω^C) of node Ω as:

$$\mathcal{P}_J(\Omega, \omega) = \max_{a_1,a_2\in\mathcal{A}_J}\ \mathbb{P}_n\left[\sum_{a_J=1}^{K_J}\hat{\mu}^{\text{AIPW}}_{J,a_J}(\mathbf{H}_J)\ \ \mathcal{I}\{A_J = g_{J,\omega,a_1,a_2}(\mathbf{H}_{\text{sub,J}}) = a_J\}\ \ \mathcal{I}(\mathbf{H}_{\text{sub,J}} \in \Omega)\right]$$

Having completed estimation of the J-th stage, we generalize estimation now for the j-th stage, each estimated in backward sequence for $j = J-1, ..., 1$. Our goal in a multi-stage setting is to estimate a DTR such that the expected long-term counterfactual outcome is optimized. When estimating the decision rule for the j-th treatment stage, we must also account for the fact that the patient was treated with the optimal treatment at all future stages. Therefore, when performing estimation for any stage prior to the last, it is necessary to calculate a stage-specific pseudo-outcome $\tilde{Y}_j$ that represents the predicted counterfactual outcome at the j-th stage contingent upon the patient receiving the optimal treatments at all future stages, $j+1, ..., J$. Mathematically this can be expressed as: $\tilde{Y}_j = \hat{E}\{Y^*(A_1, ..., A_j, g^{\text{opt}}_{j+1}, ..., g^{\text{opt}}_J)\}$, as well as in recursive form, $\tilde{Y}_j = \hat{E}\{\tilde{Y}_{j+1}|A_{j+1} = g^{\text{opt}}_{j+1}(\mathbf{H}_{\text{sub,j+1}}), \mathbf{H}_{\text{sub,j+1}}\}$. Denoting $\hat{E}(\tilde{Y}_j|A_j = a_j, \mathbf{H}_{\text{sub,j}})$ as $\tilde{\mu}_{j,a_j}(\mathbf{H}_{\text{sub,j}})$, we can express the j-th stage pseudo-outcome as $\tilde{Y}_j = \tilde{\mu}_{j+1,g^{\text{opt}}_{j+1}}(\mathbf{H}_{\text{sub,j+1}})$. Similar to the delineation above, under the assumptions of consistency, positivity, and NUCA, we express the optimal decision rule at the j-th stage as a function of the pseudo-outcome as follows:

$$g^{\text{opt}}_j(\mathbf{H}_{\text{sub}}) = \text{argmax}_{g_j\in\mathcal{G}_{\text{sub},j}}E[\tilde{Y}_j\{g(\mathbf{H}_{\text{sub}})\}] = \text{argmax}_{g_j\in\mathcal{G}_{\text{sub},j}}E_{\mathbf{H}_j}\left[\sum_{a_j=1}^{K_j}\tilde{\mu}_{j+1,a_{j+1}}(\mathbf{H}_{\text{sub,j}})\mathcal{I}\{A_j = g_j(\mathbf{H}_{\text{sub,j}}) = a_j\}\right]$$

Defining $\tilde{\mu}^{\text{AIPW}}_{j,a_j}(\mathbf{H}_j)$ as:

$$\tilde{\mu}^{\text{AIPW}}_{j,a_j}(\mathbf{H}_j) = \frac{\mathcal{I}(A_j = a_j)}{\hat{\pi}_{j,a_j}(\mathbf{H}_j)}\tilde{Y}_j + \left\{1 - \frac{\mathcal{I}(A_j = a_j)}{\hat{\pi}_{j,a_j}(\mathbf{H}_j)}\right\}\tilde{\mu}_{j,a_j}(\mathbf{H}_j)$$

where $\tilde{\mu}_{j,a_j}(\mathbf{H}_j)$ is $E(\tilde{Y}_j|A_j = a_j, \mathbf{H}_j)$, then the ReST-L purity measure used at the j-th treatment stage is:

$$\mathcal{P}_j(\Omega, \omega) = \max_{a_1,a_2\in\mathcal{A}_j}\ \mathbb{P}_n\left[\sum_{a_j=1}^{K_j}\tilde{\mu}^{\text{AIPW}}_{j,a_j}(\mathbf{H}_j)\ \ \mathcal{I}\{A_j = g_{j,\omega,a_1,a_2}(\mathbf{H}_{\text{sub,j}}) = a_j\}\ \ \mathcal{I}(\mathbf{H}_{\text{sub,j}} \in \Omega)\right]$$

3.4 | Implementation

At each stage, several user-defined inputs are needed to implement ReST-L. First, it is necessary to specify a positive value, λ_j, which is used to determine whether or not a binary split of a node identifies a meaningful difference in purity. For example, if $\mathcal{P}(\Omega_m)$ represents the purity of node Ω_m in the absence of a binary split and $\mathcal{P}(\Omega_m, \omega)$ represents the "new" purity of node Ω_m under a split defined by partition ω (and its complement ω^C), we would expect a split to occur only if a meaningful improvement in purity is achieved, i.e., $\mathcal{P}(\Omega_m, \omega) - \mathcal{P}(\Omega) > \lambda_j$. Here λ_j may be selected based on practical or clinical considerations or using cross-validation, as explained in Tao et al.[10] Additional user-defined inputs include the desired minimum size of terminal nodes and the maximum depth of the tree to be estimated, both of which may also vary by stage j. The minimum node size, $n_{0,j}$, reflects the minumum number of observations that can fall into each of the leaf nodes once a split of a parent node is made. The depth of the tree (d_j) refers to the number of times recursive splitting of the root node may occur. A smaller minimum node size and larger tree depth result in more complex tree structures with a possible concern of overfitting whereas the converse may result in underfitting. There is an abundance of literature related to selecting optimal tuning parameters for decision-tree type estimation (e.g., Hastie et al.[16] and Boehmke & Greenwell[18]); in general the choices depend on the desired complexity of the resulting estimated stage j decision rule and often are chosen adaptively from the data. Mantovani et al.[19] suggest that optimal minimum node size for a CART type estimation ranges from 1-20 and a depth of 5 is often a good starting point.[18] Another strategy frequently employed is to grow a large tree and then prune it as needed using a cost metric,[16,18,20] for example.

We briefly summarize the set of criteria for recursive partitioning of the covariate space for each stage $j = J, J-1, ..., 1$. Refer to Tao et al.[10] for additional details. Inputs into the algorithm at the j-th stage include the purity measure $\mathcal{P}_j(\Omega_m, \omega)$; the

(pseudo)-outcomes calculated via $\hat{\mu}^{\text{AIPW}}_{J,a_J}(\mathbf{H}_J)$ and $\tilde{\mu}^{\text{AIPW}}_{j,a_j}(\mathbf{H}_j)$ for the J-th or j-th stages, respectively; the minimum cut-off level for improvement in purity λ_j; the minimum terminal node $n_{0,j}$; and the maximum tree depth d_j. Beginning with the root node at the j-th stage, a series of recursive, binary splits of the covariate space $\mathbf{H}_{\text{sub},j}$ are made at the level of each node Ω_m, where the split is identified by ω, if the following criteria are met:

1. The node Ω_m resides at a shallower depth than the maximum, pre-specified tree depth d_j.
2. There are at least $2n_{0,j}$ observations in the node Ω_m and at least $n_{0,j}$ observations in each resulting child node.
3. $\mathscr{P}(\Omega_m, \omega) - \mathscr{P}(\Omega_m) > \lambda_j$, where $\mathscr{P}(\Omega_m)$ refers to the purity in the absence of a split.

If these criteria are met, we compute the estimated optimal split $\hat{\omega}^{\text{opt}} = \text{argmax}_{\omega}\{\mathscr{P}_j(\Omega, \omega)\}$. Recursive partitioning continues for the j-th stage across each node until at least one of the criteria is not met, at which point the node becomes a terminal node. Once all nodes within the j-th stage estimation become terminal, estimation for the j-th stage ends. The optimal j-stage decision rule is then determined by the partition of the covariate space at the j-th stage, with each partition being assigned the optimal treatment that maximizes the mean counterfactual (pseudo)-outcome. Estimation continues backward through all stages from the final stage J to stage 1.

4 | SIMULATION STUDIES

4.1 | Two-Stage Simulation to Evaluate the Bias of a Naive Implementation of T-RL

This first simulation demonstrates the need of our proposed method. As introduced previously, the ReST-L purity measure is constructed using the full set of covariates, $\mathbf{H}$, which are used in the AIPW estimator of the counterfactual mean outcome, but only a restricted subset of covariates are considered as candidate tailoring variables when constructing the decision tree. Given that the full set of variables can be reduced to a smaller set of variables, i.e., $\mathbf{H}_{\text{sub}}$, one may be tempted to input only those variables in $\mathbf{H}_{\text{sub}}$ into the T-RL algorithm, i.e., to estimate the AIPW estimator of the counterfactual mean outcome using only variables in $\mathbf{H}_{\text{sub}}$. We refer to this method as "Naive T-RL".

Assuming a two-stage DTR with three treatment options per stage and a sample size of $n = 1000$, we generate independent observations under varying conditions, including different levels of covariate correlations (ρ), number of variables in the full covariate history $\mathbf{H}$ and the subset $\mathbf{H}_{\text{sub}}$, and with both underlying tree-type and nontree-type DTRs. The full data generation mechanism and relevant analysis assumptions are described in detail in Supplemental Content Section 1.4. In summary, we assume that only variables in $\mathbf{H}_{\text{sub}}$ may be included in an estimated optimal DTR, but that variables from either $\mathbf{H}_{\text{sub}}$ or $\mathbf{H}^C_{\text{sub}}$ may define the intermediate outcomes and the treatment assignment mechanisms. Under optimal treatment allocation, $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 5.4$.

Results for this simulation study are presented in Table 1 . It can easily be seen that, under all data generation settings, Naive T-RL will generate a substantial bias in its estimate of the counterfactual mean outcome and a substantially lower percentage of observations correctly classified to their optimal two-stage treatment regime than ReST-L. For a tree-type DTR with 20 covariates and a correlation of $\rho = 0.2$, for example, we observe a relative bias in estimation of the optimal counterfactual mean outcome of 15.6% for Naive T-RL compared with 5.5% for ReST-L. The corresponding percentage of observations in the test set ($N_{\text{test}} = 1000$) that were correctly classified to their optimal treatment assignment for Naive T-RL and ReST-L are 57.5% and 85.3%, respectively. Refer to Supplemental Content Section 1.4 for additional simulation results reflecting this data generation setting.

4.2 | Single Stage Simulation to Evaluate Relative Performance of ReST-L

We evaluate the relative performance of ReST-L in a single stage setting with three treatment options. Parameters varied across this simulation study include the sample size, the number of covariates in $\mathbf{H}$ and $\mathbf{H}_{\text{sub}}$, the correlation used to generate the correlation matrix for covariates in $\mathbf{H}$, and the true, underlying structure of the decision rule (i.e., tree- or nontree-type). Data is generated assuming independent observations. Covariate data with dimension n x $|\mathbf{H}|$ are generated using the multivariate normal distribution with a mean of $\mathbf{0}_{|\mathbf{H}|}$ and an autoregressive (AR1) correlation structure with specified ρ and mimicking the specific correlation structure used for the simulation in Section 4.1. [Supplemental simulations using a simple exchangeable correlation structure with $\rho = 0.2$ revealed similar results to those presented herein (results not shown).] The actual treatment

received, A, is randomly generated from the multinomial distribution with probabilities π_0, π_1, π_2 where $\pi_0 = 1 - \pi_1 - \pi_2$, $\pi_1 = \exp(0.5X_{C1}+0.5X_1)/[1+\exp(0.5X_{C1}+0.5X_1)+\exp(0.5X_{C2}-0.5X_1)]$ and $\pi_2 = \exp(0.5X_{C2}-0.5X_1)/[1+\exp(0.5X_{C1}+0.5X_1) + \exp(0.5X_{C2} - 0.5X_1)]$, where X_{C1}, X_{C2} represent the first two covariates in $\mathbf{H}^C_{\text{sub}}$, i.e., confounding variables not considered as candidate tailoring variables. The outcome $Y = \exp\{1.5+0.3X_{C1}-|1.5X_1-2|\cdot(A-g^{\text{opt}})^2\}+\epsilon$, where $\epsilon \sim N(0,1)$. The true, underlying tree-type decision rule is defined as follows: If $X_1 > -1$ & $X_2 > 0.5$, then $g^{\text{opt}} = 2$; if $X_1 > -1$ & $-0.5 < X_2 < 0.5$, then $g^{\text{opt}} = 1$; otherwise, $g^{\text{opt}} = 0$. The nontree-type decision rule is defined as: $g^{\text{opt}} = \mathcal{I}\{\log_2(|X_1|+1) \leq 2$ & $X_2 < 0.25\} + \mathcal{I}\{X_2^2 \leq 0.5\}$. Importantly, the outcome and actual treatment assignment are defined using variables in both $\mathbf{H}_{\text{sub}}$ and $\mathbf{H}^C_{\text{sub}}$. The optimal decision rule, based on the methodologic assumptions of ReST-L, includes only variables in $\mathbf{H}_{\text{sub}}$. Under optimal treatment allocation $\hat{E}[Y^*\{g^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 4.7$.

We compare the estimated performance of ReST-L with five competing methods: tree-based reinforcement learning (T-RL), standard Q-Learning using linear modeling (Q-L), restricted linear Q-Learning (Q-L-R), Q-Learning using nonparametric modeling (Q-NP), and restricted nonparametric Q-Learning (Q-NP-R). With the exception of T-RL, which represents the unrestricted counterpart to ReST-L, we restrict our comparisons to Q-Learning methods because these are the only existing methods to our knowledge that can accommodate a subset of variables in the estimated treatment regime. For both ReST-L and T-RL we assume that there is an additive linear relationship between the outcome Y and covariate or treatment history that includes all observed covariates, as well as a treatment-interaction with either all observed covariates (T-RL) or with a subset of candidate tailoring variables (ReST-L). We further assume that the propensity model used in ReST-L and T-RL is correctly specified. (Performance results under an incorrectly-specified propensity model are presented for a two-stage simulation in the Supplemental Content Section 1.1.) Restricted Q-Learning methods are modifications of the standard Q-Learning models such that only variables in $\mathbf{H}_{\text{sub}}$ are considered as possible candidate tailoring variables (i.e., treatment interactions), which differs from standard Q-Learning in which all variables in $\mathbf{H}$ are possible treatment tailoring variables. Linear Q-Learning assumes a linear relationship between the covariates and the outcome. Nonparametric Q-Learning methods allow a more flexible relationship for the Q-functions, estimated using random forests (`randomForest` in `R`[21]). Performance is evaluated using two metrics: 1) the optimal treatment regime using data from the training set with sample size n and use a test set ($N_{\text{test}} = 1000$) to determine the percentage of observations correctly classified to their optimal treatment, *%opt*; and 2) $E[Y^*\{\hat{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$, the expected counterfactual outcome had everyone in the patient population been treated optimally based on the estimated optimal regime, estimated using the test set. For each design setting, we tabulate the median and interquartile range (IQR) of *%opt* and $\hat{E}[Y^*\{\hat{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ across all $B = 500$ Monte Carlo iterations.

Estimated performance for tree-type and nontree-type decision rules are displayed in Table 2 . As observed in the tabulated results, ReST-L selects the optimal treatment decision rule well when the underlying decision rule is either tree-type or nontree-type. Across all data generating settings, for a tree-type decision rule the percent of observations from the test set that are correctly classified to their optimal treatment ranges from about 85% for smaller sample sizes and fewer variables in $\mathbf{H}$ and $\mathbf{H}_{\text{sub}}$ to more than 95% for larger sample sizes (and a correspondingly larger number of variables in $\mathbf{H}$). ReST-L performance improves as the sample size increases, as expected; for example, refer to results for sample sizes of $n = 500$ and $n = 750$ when $|\mathbf{H}| = 100$. For the same sample size and number of variables in the covariate history $\mathbf{H}$, performance improves as the proportion of variables in $\mathbf{H}_{\text{sub}}$ relative to $\mathbf{H}$ decreases. For example, for 50 covariates in $\mathbf{H}$, a sample size of $n = 300$ and a correlation $\rho = 0.2$, estimated performance improves from 86.5% to 90.0% correct classification when the number of variables in $\mathbf{H}_{\text{sub}}$ decreases from 35 to 10. ReST-L performance in estimating the optimal decision rule is similar across different degrees of correlation among covariates when all other parameters are held constant. Finally, we observe that the variability for ReST-L in estimating the optimal regime increases when the true, underlying decision rule is nontree-type, and is generally higher with a smaller sample size or as the proportion of variables in $\mathbf{H}_{\text{sub}}$ increases. Furthermore, ReST-L estimates the empirical counterfactual mean outcome under the optimal treatment regime, $\hat{E}[Y^*\{\hat{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$, with a high degree of accuracy and relatively low variability across Monte Carlo iterations, particularly as the sample size increases. As an example, assuming a tree-type decision rule, $|\mathbf{H}| = 100$, and $n = 750$, ReST-L estimates the counterfactual mean under the estimated optimal treatment assignment to be 4.6, which is very close to the true empirical counterfactual mean of 4.7.

ReST-L consistently performs better than all other methods across all one-stage data generating settings presented – for both tree- and nontree-type decision rules. For estimation with an underlying tree-type decision rule, the variability of ReST-L in estimating the optimal regime is smaller than that of T-RL and similar to restricted nonparametric Q-Learning. For a nontree-type decision rule, variability of ReST-L in estimating *%opt* is larger than that of restricted nonparametric Q-Learning, but generally remains smaller than that of T-RL overall. Across all simulation settings, restricted Q-Learning methods perform better than their standard Q-Learning counterparts. Both linear Q-learning methods (restricted and unrestricted) perform poorly in all

scenarios whether the underlying decision rule is tree-type or nontree-type. With a larger sample size, restricted nonparametric Q-Learning does a reasonable job of estimating the optimal treatment regime for an underlying tree- and nontree-type decision rule; with $n = 500$ and $100/20$ variables in $\mathbf{H}/\mathbf{H}_{\text{sub}}$, for example, restricted nonparametric Q-Learning achieves higher than 85% correct classification.

4.3 | Two-Stage Simulation to Evaluate Relative Performance of ReST-L

We next evaluate the performance of ReST-L in a two-stage estimation setting with 3 possible treatment options per stage. It can easily be seen that random allocation of one of three treatments in each of two stages would select the optimal two-stage treatment assignment about 1 out of every 3^2 times, which is about 11% of the time. All settings for generating first stage data, including the covariate matrix $\mathbf{X}$, the treatment assignment mechanism for A_1, the intermediate outcome Y_1, and optimal treatment $g_1^{\text{opt}}(\mathbf{H}_{\text{sub}})$, are the same as those used in the single stage setting described above. The second stage treatment A_2 is randomly generated using the multinomial distribution with probabilities $\pi_{20}, \pi_{21}, \pi_{22}$ where $\pi_{20} = 1 - \pi_{21} - \pi_{22}$, $\pi_{21} = \{\exp(0.2Y_1 - 0.5)\}/[1 + \{\exp(0.2Y_1 - 0.5)\} + \{\exp(0.5X_{C2})\}]$, and $\pi_{22} = \{\exp(0.5X_{C2})\}/[1 + \{\exp(0.2Y_1 - 0.5)\} + \{\exp(0.5X_{C2})\}]$. The intermediate outcome is $Y_2 = \exp\{1.18 + 0.2X_{C2} - |1.5X_3 + 2| \cdot (A_2 - g_2^{\text{opt}})^2\} + \epsilon$, where $\epsilon \sim N(0, 1)$, and the overall outcome $Y = Y_1 + Y_2$. When a tree-type DTR is assumed, $g_2^{\text{opt}}(\mathbf{H}_{\text{sub}})$ is assigned as follows: If $X_3 > -1$ & $Y_1 > 2$, then $g_2^{\text{opt}} = 2$; if $X_3 > -1$ & $0 < Y_1 \leq 2$, then $g_2^{\text{opt}} = 1$; otherwise, $g_2^{\text{opt}} = 0$. Under an assumed nontree-type DTR: $g_2^{\text{opt}} = \mathcal{I}(|X_3| > 0.6 \;\&\; Y_1 > 0.4) + \mathcal{I}(Y_1^2 > 2.5)$. Both the intermediate outcomes and actual treatment assignments depend on variables in both $\mathbf{H}_{\text{sub}}$ and $\mathbf{H}_{\text{sub}}^C$. However, the optimal DTR are set to include only variables in $\mathbf{H}_{\text{sub}}$. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 8.0$. Similar to the single stage setting, ReST-L and T-RL assume a correctly-specified propensity model and an incorrectly-specified conditional mean model.

The performance of ReST-L and other competing methods for estimating the optimal two-stage regime with either an underlying tree-type or nontree-type DTR are displayed in Table 3 . Across all sample size and variable settings with a tree-type DTR, ReST-L does a reasonably good job of selecting the optimal treatment, with correct classification generally between 85-90%. As in a single stage estimation setting, performance improves with sample size, with an improvement in percent correct classification from 89.6% to 95.2% for sample sizes of $n = 600$ to $n = 1000$ ($\rho = 0.2$). Additionally, performance improves with fewer variables in $\mathbf{H}_{\text{sub}}$ relative to $\mathbf{H}$: The percent correct classification with 50 variables in $\mathbf{H}$ and $\rho = 0.2$ improves from 87.0% to 89.8% as the number of variables in $\mathbf{H}_{\text{sub}}$ is reduced from 35 to 10. With an underlying nontree-type DTR, larger sample sizes are needed to obtain a similar estimated correct classification rate. For example, with a sample size of $n = 600$, $\rho = 0.2$, and $|\mathbf{H}| = 100$ variables, the percent of observations correctly classified to their optimal treatment is just over 70% for the nontree-type DTR compared with nearly 90% for a tree-type DTR; however, with the same specifications but with $n = 1000$, the percent correct classification are similar for tree- and nontree-type DTRs (95.2% and 93.8%, respectively). Variability of estimation of the percent correct treatment allocation of ReST-L is lower for a tree-type DTR than for nontree-type DTR and is larger on average than that observed in a single stage setting. Finally, for ReST-L, the estimated counterfactual mean outcome is closer to the empirical mean when sample size increases; when $n = 1000$, ReST-L achieves an estimated counterfactual mean outcome of 7.8 compared to the empirical mean of 8.0.

For a two-stage, tree-type DTR, ReST-L improves more upon T-RL at lower sample sizes and when the proportion of variables in $\mathbf{H}_{\text{sub}}$ relative to $\mathbf{H}$ decreases. With larger sample sizes, e.g., when $n = 1000$, both ReST-L and T-RL achieve more than 90% correct treatment classification although ReST-L still slightly outperforms T-RL in this case. For a nontree-type DTR, ReST-L improves upon T-RL across all settings, although in particular the benefit of ReST-L is observed with a larger number of covariates. When $n = 1000$ and $\rho = 0.2$, for example, the percent of observations in the test set that were correctly classified to their optimal treatment is 93.8% for ReST-L compared with 84.8% for T-RL. As in a single stage setting, we observe that the restricted versions of Q-Learning improve upon their unrestricted counterparts, although linear Q-Learning demonstrates poor performance across all settings, never exceeding more than 25% correct treatment classification. Restricted nonparametric Q-Learning, on the other hand, achieves good performance, nearing 90% correct classification for a tree-type DTR with a large sample size.

4.4 | Supplemental Simulation Experiments

Supplemental simulation studies were conducted to evaluate the performance of ReST-L in a variety of other scenarios. The simulation setups and results are presented in Supplemental Content Section 1. Specifically, we apply ReST-L and T-RL in a two-stage setting using an incorrectly-specified propensity model (Supplemental Content Section 1.1); or modify the data generating mechanisms to remove confounding of the treatment assignments $\mathbf{A}$ and outcomes $\mathbf{Y}$ by variables in $\mathbf{H}^C_{\text{sub}}$ (Supplemental Content Section 1.2); or modify the data generating mechanisms for $\mathbf{g}$ such that variables defining the true optimal DTR may be in $\mathbf{H}^C_{\text{sub}}$ (Supplemental Content Section 1.3); or evaluate the relative performance of ReST-L compared with other methods under stronger confounding with a binary covariate $Z \in \mathbf{H}^C_{\text{sub}}$ (Supplemental Content Section 1.4) which mimics the data generation model from Section 4.1. In Supplemental Content Section 1.5 we present simulation results involving a three-stage estimation setting in order to illustrate how these methods can apply to more than two stages.

5 | APPLICATION TO PERSONALIZE EARLY FLUID RESUSCITATION STRATEGIES IN ACUTE SEPTIS PATIENTS

Sepsis is a clinical syndrome characterized by systemic inflammation and infection and is associated with one of the highest rates of mortality among conditions commonly treated in emergency departments (EDs) and intensive care units (ICUs).[12] Due to the large degree of heterogeneity in presentation, which may include varying degrees of organ dysfunction, sepsis is a difficult condition to diagnose and even more difficult to successfully treat. Sepsis is routinely treated using fluid resuscitation, antibiotics, and may also include treatment with vasopressors, mechanical ventilation, and others. The established clinical guidelines for treating sepsis, known as the "Surviving Sepsis Campaign",[13] strongly recommends that resuscitation of at least 30 mL/kg of IV fluid be given within the first 3 hours. However, this recommendation is given with a stated "low quality of evidence" due to the fact that results across studies have been inconsistent with indirect evidence, imprecise results, and a likelihood of bias.

Due to the paucity of strong evidence as to the most beneficial fluid resuscitation strategy in the early hours of treatment, we estimate an optimal two-stage DTR in septic patients admitted to an ICU after presenting to the ED. Using electronic medical record and administrative data from the Medical Information Mart for Intensive Care III (MIMIC-III),[22,23,24] a retrospectively-collected and freely-available database accessible through PhysioNet[25] that contains de-identified and anonymized data for patients treated in an ICU at a tertiary care medical facility, we evaluate whether treatment with fluid restrictive or fluid liberal strategies can be further tailored in order to minimize organ dysfunction scores. Stage 1 treatment was defined as either a fluid restrictive (< 30 mL/kg) or a fluid liberal (≥ 30 mL/kg) strategy within the first three hours after admission to the ICU. Baseline covariates considered as candidate tailoring variables for a first-stage treatment rule included age, Elixhauser comorbidity score,[26,27] and BMI. Covariates excluded from consideration as a possible Stage 1 tailoring variable were racial/ethnic identity, gender, the ICU unit in which the patient was first treated, and the time of year in which the patient was treated. Stage 2 treatment is defined as either a fluid restrictive (< 30 mL/kg) or a fluid liberal (≥ 30 mL/kg) strategy between 3-24 hours after ICU admission. Intermediate variables collected prior to Stage 2 treatment included indicators of treatment with mechanical ventilation and vasopressors within the first three-hour time period, as well as the patient's SOFA score evaluated at three hours post-admission. The final outcome of interest is the Sequential Organ Failure Assessment (SOFA)[28] score evaluated at 24 hours post-admission. Refer to Supplemental Content for specific cohort eligibility and additional analysis details.

Seven hundred eight (708) patients were included in the analysis cohort. The average patient was a 68 year old, overweight, white (76%), male (54%) with 0 reported Elixhauser comorbidities (Table 4). The median length of hospital stay was 8.9 days with an interquartile range (IQR) of 5.7-15.8. The median fluid input received within 0-3 hours and 3-24 hours post-admission is 38.0 mL/kg (IQR: 20.5-57.8) and 16.8 mL/kg (IQR: 0.0-44.1), respectively. Summary statistics stratified by treatment stage (i.e., 0-3 hours and 3-24 hours post-ICU admission) demonstrate covariate imbalance for age, gender, and race/ethnicity across fluid resuscitation strategies in the first treatment stage, and for weight, BMI, and the use of mechanical ventilation and vasopressors across fluid resuscitation strategies for both stages, suggesting that confounding is an issue that must be addressed in our analysis in order to make causal inference.

As can be observed in Figure 1, it is recommended that all patients with a BMI classification of either "normal" or "overweight" (i.e., 18.5-29.9 kg/m^2) should receive liberal fluid resuscitation (≥ 30 mL/kg) within the first three hours following admission to the ICU for treatment of acute emergent sepsis, but restrictive fluid resuscitation if the patient is classified as either "obese" (> 30.0 kg/m^2) or "underweight" (< 18.5 kg/m^2). Within 3-24 hours post-admission, all patients should receive a restrictive

fluid resuscitation strategy in order to minimize the SOFA score at 24 hours. This estimated DTR reflects a SOFA improvement of roughly 0.1.

Although the question of how to optimally treat septic patients is complex and multi-faceted, we applied a robust and flexible causal method with interpretable results to determine whether tailoring of fluid resuscitation strategies at each of two stages within the first 24 hours after ICU admission can be used to improve outcomes overall. In contrast to the Surviving Sepsis Campaign, which recommends early liberal fluid resuscitation for all septic patients, our estimated optimal DTR assigns only patients in an average BMI class (i.e., either normal or overweight) to early liberal fluid resuscitation within the first three hours following admission. Given that the volume of fluids administered in resuscitation strateiges is defined using the patient's weight, our results suggest that additional investigation into the application of weight-based dosing strategies to treat acute septic patients may be needed.

6 | DISCUSSION

Personalized medicine reflects a goal of providing the right treatment to the right person at the right time. ReST-L provides a flexible, data-driven approach grounded in causal inference for estimating an interpretable, optimal multi-stage DTR using observational data when only a subset of covariates, based on clinical or other knowledge, should be considered as candidate tailoring variables. Importantly, ReST-L addresses a clinical scenario that has not yet been addressed to our knowledge in the literature for tree-based, optimal DTR estimation. We have shown that there is an improvement over other estimation methods when a clinically-meaningful or ethical treatment decision should be made without certain variables and, given that ReST-L reduces to T-RL when the full set of covariates are considered, this provides an important extension of previous work. ReST-L utilizes a purity measure that is based upon a consistent and doubly-robust estimator of the counterfactual mean outcome under a sub-tree regime when either the propensity model or the conditional mean model are correctly specified, resulting in a causal estimator with double protections against model misspecifications. We demonstrate that ReST-L can estimate the optimal, multi-stage DTR in the presence of a moderately large degree of covariates and we base simulation studies on a reasonably complex relationship that is intended to be reflective of data generating mechanisms that may be seen in the real world.

Our results reflect a small number of possible data generating scenarios and it is likely that performance estimates would change under different simulation settings. We do, however, conduct simulation studies under varying assumptions and believe these comprehensive results provide a solid understanding of ReST-L performance. For estimations in a two-stage setting, we observe a high degree of variability in the estimated percentage of observations correctly classified to their optimal treatment using ReST-L. While the variability is much lower than the variability observed in T-RL, it is much larger than that estimated using restricted Q-Learning with nonparametric modeling assumptions. However, the median estimated performance is also consistently higher for ReST-L in a two-stage setting than it is for restricted nonparametric Q-Learning, suggesting that a trade of higher variability for higher estimated performance could be warranted. As was shown in the supplemental content, ReST-L also presents a reasonable solution for estimating DTRs with three or more stages. In practice, however, we do find that it is often difficult to obtain sufficient data to answer a well-constructed three-stage research question. Therefore, eagerness in these instances should be tempered with pragmatism. Finally, in our simulation studies we assume that the conditional mean models are incorrectly specified. Although this is useful in order to provide an understanding of performance as an "out of the box" solution for optimal DTR estimation, model selection and diagnostics can be used to select either the propensity or the conditional mean model, or both. This was not explored, but this may be considered in data applications and/or in future research.

There is one final point we would like to point out that we believe is of interest to readers. ReST-L is a method for DTR estimation, but does not at this time provide a solution for inference. Given that our estimand is a multi-stage treatment regime and not a point estimate per se, a confidence interval would be impractical. We believe a primary question of interest, however, is whether a particular estimated DTR provides a statistically-significant improvement in outcomes compared with another (nested) DTR. At this time, ReST-L addresses "meaningful improvements" in outcome through the use of the tree-building tuning parameter λ. In the future, however, we could conceivably approach this need by estimating confidence intervals for the expected outcome under a specific regime or, perhaps, by using a hypothesis testing approach at the level of the binary covariate splits. Regardless, we believe this is an interesting and important question for future work.

ACKNOWLEDGEMENTS

The authors would like to thank the MIMIC-III and Physionet groups for providing the dataset and supporting materials and Dr. Juan Luis Marquez for the helpful discussions related to the data application.

DATA AVAILABILITY STATEMENT

The Medical Information Mart for Intensive Care III (MIMIC-III) data are available through Physionet (mimic.physionet.org). Public access to code used to perform simulations and data application is available at https://github.com/kellyspeth/ReSTL.

ORCID

Kelly Speth, https://orcid.org/0000-0001-8045-8406

REFERENCES

References

1. Chakraborty B, Moodie E. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine. Statistics for Biology and Health.* New York, NY: Springer . 2013.
2. Murphy S. Optimal Dynamic Treatment Regimes. *J R Stat Soc Series B Stat Methodol* 2003; 65(2): 331-355.
3. Tsiatis A, Davidian M, Holloway S, Laber E. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine. Monographs on Statistics and Applied Probability*. Boca Raton, FL: CRC Press . 2020.
4. Laber E, Zhao Y. Tree-based methods for individualized treatment regimes. *Biometrika* 2015; 102(3): 501-514.
5. Zhao Y, Zeng D, Rush J, Kosorok M. Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *J Am Stat Assoc* 2012; 107(449): 1106-1118.
6. Zhao Y, Zeng D, Laber E, Kosorok M. New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes. *J Am Stat Assoc* 2015; 110(510): 583-598.
7. Zhang B, Tsiatis A, Davidian M, Zhang M, Laber E. Estimating optimal treatment regimes from a classification perspective. *Stat* 2012; 1: 103-114.
8. Zhang B, Tsiatis A, Laber E, Davidian M. A Robust Method for Estimating Optimal Treatment Regimes. *Biometrics* 2012; 68(4): 1010-1018.
9. Zhang Y, Laber E, Davidian M, Tsiatis A. Interpretable Dynamic Treatment Regimes. *J Am Stat Assoc* 2018; 113(524): 1541-1549.
10. Tao Y, Wang L, Almirall D. Tree-based Reinforcement Learning for Estimating Optimal Dynamic Treatment Regimes. *Ann Appl Stat* 2018; 12(3): 1914-1938.
11. Breiman L, Freidman J, Olshen R, Stone C. *Classification and regression trees*. Belmont, CA: Wadsworth . 1984.
12. Marino P. *Marino's The ICU Book, 4th ed.* Philadelphia, PA: Wolters Kluver Health/Lippincott Williams and Wilkins . 2014.
13. Rhodes A, Evans L, Waleed A, et al. . Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. *Intensive Care Med* 2017; 45(3): 486-552.

14. Rubin D. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *J Educ Psychol* 1974; 66: 688-701.

15. Tao Y, Wang L. Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. *Biometrics* 2017; 73: 145-155.

16. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed.* New York, NY: Springer . 2009.

17. Bather J. *Decision Theory: An Introduction to Dynamnic Programming and Sequential Decisions*. New York, NY: Wiley . 2000.

18. Boehmke B, Greenwell B. *Decision Trees In: Hands-On Machine Learning with R* . 2020.

19. Mantovani R, Horvath T, Cerri R, Junior S, Vanschoren J, de Carvalho A. An empirical study on hyperparameter tuning of decision trees. *arXiv* 2019; 2.

20. Therneau T, Atkinson E, Mayo Foundation . An Introduction to Recursive Partitioning Using the RPART Routines. *CRAN R Network* 2019.

21. R Core Team . *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing . 2018.

22. Johnson A, Pollard T, Shen L, et al. . MIMIC-III, a freely accessible critical care database. *Sci data* 2016; 3: 160035.

23. Johnson A, Stone D, Celi L, Pollard T. The MIMIC Code Repository: enabling reproducibility in critical care research. *J Am Med Inform Assoc* 2017.

24. Pollard TJ, Johnson AE. The MIMIC-III Clinical Database. http://dx.doi.org/10.13026/C2XW26; 2016

25. Goldberger A, Amaral L, Glass L, Hausdorff J, et al. . PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 2000; 101(23): e215–e220.

26. Elixhauser A, Steiner C, Harris D, et al. . Comorbidity measures for use with administrative data. *Med Care* 1998; 36: 8-27.

27. van Walraven C, Austin P, Jennings A, et al. . A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med Care* 2009; 47: 626-633.

28. Vincent J, Moreno R, Takala J, Willatts S, et al. . The SOFA (Sequential Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Med* 1996; 22(7): 707-710.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supplemental Information section at the end of the article.

TABLE 1 Simulation results comparing Naive T-RL and ReST-L in estimating an optimal two-stage dynamic treatment regime (DTR) with three treatments per stage and a high degree of confounding, demonstrating the need for our new method, ReST-L. Medians (and IQRs) of $\hat{E}[Y^*\{\hat{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ and % *opt*, as well as absolute and relative bias, are presented. $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; Naive T-RL = Naive Tree-based Reinforcement Learning; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment; IQR = interquartile range; Abs = Absolute Bias; Rel % = relative percent bias; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a DTR estimated using the applicable method.

$\lvert\mathbf{H}\rvert/\lvert\mathbf{H}\text{sub}\rvert$	ρ	Naive T-RL			ReST-L		
		$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)	Abs Bias (Rel %)	%opt (IQR)	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)	Abs Bias (Rel %)	%opt (IQR)
Tree-type DTR							
20/7	0	4.542 (0.588)	0.869 (16.1)	55.0 (20.8)	5.133 (0.269)	0.278 (5.1)	86.5 (10.6)
20/7	0.2	4.569 (0.552)	0.842 (15.6)	57.5 (20.4)	5.114 (0.256)	0.297 (5.5)	85.3 (10.2)
20/7	0.6	4.569 (0.490)	0.842 (15.5)	55.5 (20.1)	5.123 (0.290)	0.288 (5.3)	85.4 (10.5)
50/10	0	4.491 (0.525)	0.920 (17.0)	54.1 (18.6)	5.114 (0.251)	0.297 (5.5)	85.4 (11.4)
50/10	0.2	4.589 (0.523)	0.822 (15.2)	57.2 (21.6)	5.106 (0.265)	0.305 (5.6)	85.2 (12.0)
50/10	0.6	4.525 (0.538)	0.886 (16.4)	54.3 (19.9)	5.093 (0.292)	0.318 (5.9)	84.4 (12.1)
50/35	0	4.489 (0.558)	0.922 (17.0)	53.8 (16.9)	5.056 (0.290)	0.355 (6.6)	82.3 (13.6)
50/35	0.2	4.485 (0.553)	0.926 (17.1)	54.0 (19.0)	5.049 (0.298)	0.362 (6.7)	82.3 (13.2)
50/35	0.6	4.483 (0.541)	0.928 (17.2)	52.3 (20.2)	5.049 (0.267)	0.362 (6.7)	81.1 (11.9)
100/20	0	4.475 (0.587)	0.936 (17.3)	53.4 (19.0)	5.030 (0.319)	0.381 (7.0)	82.0 (13.5)
100/20	0.2	4.560 (0.543)	0.851 (15.7)	56.2 (19.7)	5.030 (0.306)	0.381 (7.0)	82.7 (11.6)
100/20	0.6	4.519 (0.510)	0.892 (16.5)	54.9 (20.1)	5.030 (0.328)	0.381 (7.0)	81.6 (12.9)
Nontree-type DTR							
20/7	0	4.797 (0.763)	0.614 (11.3)	66.3 (26.3)	5.132 (0.373)	0.279 (5.2)	82.8 (17.2)
20/7	0.2	4.656 (0.701)	0.755 (14.0)	64.2 (25.7)	5.107 (0.350)	0.304 (5.6)	82.3 (15.9)
20/7	0.6	4.612 (0.666)	0.799 (14.8)	63.1 (19.1)	5.118 (0.364)	0.293 (5.4)	81.4 (15.2)
50/10	0	4.740 (0.737)	0.671 (12.4)	65.7 (25.1)	5.092 (0.428)	0.319 (5.9)	81.6 (20.4)
50/10	0.2	4.646 (0.678)	0.765 (14.1)	64.5 (24.0)	5.118 (0.407)	0.293 (5.4)	82.7 (18.4)
50/10	0.6	4.506 (0.593)	0.905 (16.7)	60.4 (16.2)	5.047 (0.461)	0.364 (6.7)	80.0 (21.6)
50/35	0	4.708 (0.715)	0.703 (13.0)	64.0 (25.1)	5.027 (0.520)	0.384 (7.1)	77.6 (20.5)
50/35	0.2	4.659 (0.715)	0.752 (13.9)	64.3 (24.4)	5.004 (0.466)	0.407 (7.5)	79.4 (19.4)
50/35	0.6	4.465 (0.482)	0.946 (17.5)	58.0 (14.6)	4.971 (0.564)	0.440 (8.1)	76.6 (22.6)
100/20	0	4.706 (0.712)	0.705 (13.0)	65.4 (24.3)	5.002 (0.575)	0.409 (7.6)	78.6 (22.7)
100/20	0.2	4.594 (0.663)	0.817 (15.1)	63.3 (21.6)	5.013 (0.537)	0398 (7.4)	79.0 (22.4)
100/20	0.6	4.479 (0.540)	0.932 (17.2)	59.4 (16.3)	4.988 (0.507)	0.423 (7.8)	76.7 (21.3)

TABLE 2 Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal one-stage decision rule with 3 possible treatments based on an underlying, tree-type (top panel) and nontree-type (bottom panel) decision rule. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment rule estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment.

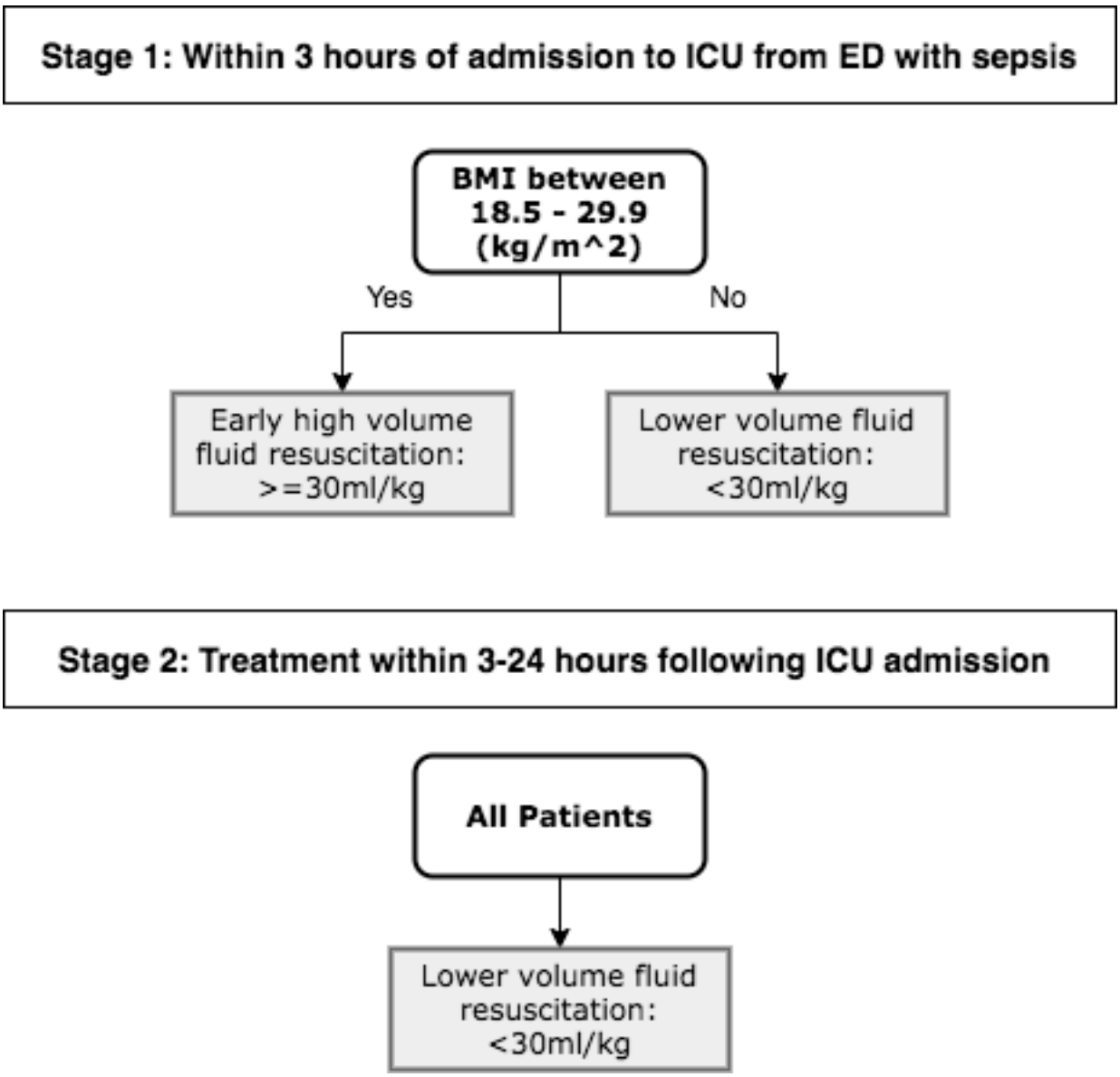
n	$\lvert\mathbf{H}\rvert/\lvert\mathbf{H}\text{sub}\rvert$	rho	ReST-L		T-RL		Q-L-R		Q-L		Q-NP-R		Q-NP	
			%opt	$\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$	%opt	$\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$	%opt	$\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$	%opt	$\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$	%opt	$\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$	%opt	$\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$
Tree-type DTR														
750	100/20	0.2	97.0 (4.4)	4.6 (0.2)	94.6 (5.9)	4.5 (0.2)	62.3 (2.3)	3.4 (0.1)	56.4 (2.3)	3.1 (0.1)	93.8 (3.9)	4.5 (0.1)	84.4 (7.0)	4.2 (0.3)
750	100/20	0.6	96.6 (4.5)	4.6 (0.2)	94.5 (5.3)	4.5 (0.2)	62.3 (2.2)	3.4 (0.1)	56.4 (2.9)	3.1 (0.1)	92.4 (3.9)	4.5 (0.1)	82.8 (5.9)	4.1 (0.2)
500	100/20	0.2	93.6 (5.8)	4.5 (0.2)	84.8 (13.0)	4.1 (0.4)	61.4 (2.4)	3.3 (0.1)	52.8 (3.1)	2.9 (0.1)	88.7 (5.4)	4.3 (0.2)	75.8 (7.4)	3.8 (0.3)
500	100/20	0.6	93.3 (5.7)	4.5 (0.2)	86.2 (9.6)	4.2 (0.4)	61.1 (2.3)	3.3 (0.1)	52.6 (3.2)	2.9 (0.2)	86.5 (5.4)	4.2 (0.2)	72.6 (7.1)	3.7 (0.3)
300	50/35	0.2	86.5 (11.1)	4.2 (0.4)	81.6 (15.6)	4.0 (0.6)	56.8 (3.2)	3.1 (0.2)	54.2 (3.5)	3.0 (0.2)	72.8 (9.0)	3.7 (0.3)	66.0 (9.7)	3.5 (0.4)
300	50/35	0.6	85.9 (10.5)	4.2 (0.4)	82.7 (13.5)	4.1 (0.5)	56.9 (3.2)	3.1 (0.1)	54.3 (3.6)	3.0 (0.2)	70.9 (7.5)	3.7 (0.3)	63.4 (8.0)	3.4 (0.3)
300	50/10	0.2	90.0 (6.9)	4.3 (0.3)	82.4 (14.2)	4.0 (0.5)	61.8 (2.4)	3.3 (0.1)	54.4 (3.2)	3.0 (0.2)	83.4 (6.8)	4.1 (0.2)	66.1 (8.9)	3.5 (0.3)
300	50/10	0.6	90.4 (7.3)	4.3 (0.3)	82.5 (14.1)	4.1 (0.5)	61.6 (2.4)	3.3 (0.1)	54.3 (3.3)	3.0 (0.2)	80.6 (6.9)	4.0 (0.2)	62.9 (8.4)	3.4 (0.3)
200	20/7	0.2	86.4 (8.5)	4.2 (0.3)	83.8 (12.6)	4.0 (0.4)	61.9 (2.7)	3.3 (0.1)	58.2 (3.4)	3.1 (0.2)	78.5 (7.9)	4.0 (0.3)	64.3 (9.7)	3.4 (0.4)
200	20/7	0.6	86.3 (6.8)	4.2 (0.3)	84.4 (10.7)	4.1 (0.4)	61.7 (2.7)	3.3 (0.1)	58.1 (3.5)	3.2 (0.2)	76.2 (7.8)	3.9 (0.3)	62.5 (8.0)	3.4 (0.3)
Nonree-based DTR														
750	100/20	0.2	98.6 (2.1)	4.6 (0.1)	97.8 (6.7)	4.6 (0.2)	50.2 (3.9)	2.8 (0.2)	40.2 (2.8)	2.4 (0.1)	93.0 (3.6)	4.5 (0.1)	85.0 (8.9)	4.2 (0.3)
750	100/20	0.6	98.4 (2.3)	4.6 (0.1)	97.7 (8.5)	4.6 (0.3)	50.3 (3.9)	2.9 (0.2)	40.4 (2.9)	2.5 (0.1)	91.5 (4.1)	4.5 (0.1)	77.2 (10.1)	3.9 (0.4)
500	100/20	0.2	95.8 (14.1)	4.5 (0.5)	78.6 (25.0)	4.0 (0.9)	47.3 (4.5)	2.7 (0.2)	37.4 (3.1)	2.3 (0.1)	87.1 (5.5)	4.3 (0.2)	70.5 (11.1)	3.7 (0.4)
500	100/20	0.6	94.4 (18.9)	4.5 (0.6)	68.8 (24.8)	3.6 (0.8)	47.6 (5.0)	2.8 (0.2)	37.7 (3.1)	2.3 (0.2)	85.5 (5.4)	4.3 (0.2)	66.7 (10.4)	3.5 (0.4)
300	50/35	0.2	83.9 (29.9)	4.1 (1.0)	78.0 (26.5)	3.9 (1.0)	40.8 (4.3)	2.4 (0.2)	38.7 (4.0)	2.3 (0.2)	66.7 (8.8)	3.6 (0.3)	61.3 (9.8)	3.3 (0.4)
300	50/35	0.6	83.4 (29.7)	4.1 (1.0)	77.8 (25.7)	3.9 (0.9)	41.2 (4.1)	2.5 (0.2)	39.0 (3.7)	2.4 (0.2)	67.9 (8.0)	3.6 (0.3)	58.6 (8.6)	3.3 (0.3)
300	50/10	0.2	94.6 (16.9)	4.5 (0.6)	78.6 (26.6)	3.9 (1.0)	49.0 (5.6)	2.8 (0.2)	38.9 (3.6)	2.3 (0.2)	79.6 (6.9)	4.0 (0.2)	60.7 (9.3)	3.3 (0.3)
300	50/10	0.6	94.0 (16.7)	4.5 (0.6)	76.9 (25.9)	3.9 (0.9)	49.0 (5.4)	2.8 (0.3)	38.9 (3.6)	2.4 (0.2)	78.2 (7.0)	4.0 (0.2)	58.0 (8.9)	3.2 (0.3)
200	20/7	0.2	93.1 (31.6)	4.4 (1.1)	82.9 (31.7)	4.1 (1.1)	48.9 (7.0)	2.8 (0.3)	42.0 (4.7)	2.5 (0.2)	74.2 (8.3)	3.8 (0.3)	60.5 (9.4)	3.3 (0.4)
200	20/7	0.6	92.2 (32.6)	4.4 (1.1)	82.2 (32.3)	4.1 (1.1)	49.0 (6.3)	2.8 (0.3)	42.4 (4.8)	2.6 (0.2)	72.8 (7.1)	3.8 (0.3)	56.5 (8.5)	3.2 (0.4)

TABLE 3 Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments per stage and based on an underlying, tree-type DTR (top panel) and nontree-type DTR (bottom panel). n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment regime estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment.

n	$\lvert\mathbf{H}\rvert/\lvert\mathbf{H}\text{sub}\rvert$	rho	ReST-L		T-RL		Q-L-R		Q-L		Q-NP-R		Q-NP	
			%opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	%opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	%opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	%opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	%opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	%opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$
Tree-type DTR														
1000	100/20	0.2	95.2 (5.5)	7.8 (0.2)	92.5 (7.3)	7.8 (0.3)	51.0 (2.1)	5.9 (0.1)	45.1 (2.6)	5.5 (0.2)	88.9 (4.8)	7.7 (0.2)	75.0 (8.6)	7.1 (0.4)
1000	100/20	0.6	93.8 (5.9)	7.8 (0.2)	91.8 (7.2)	7.7 (0.3)	50.9 (2.3)	5.9 (0.2)	44.8 (2.7)	5.6 (0.2)	86.9 (4.7)	7.6 (0.2)	73.1 (6.6)	7.1 (0.3)
600	100/20	0.2	89.6 (10.4)	7.7 (0.3)	77.4 (17.1)	7.2 (0.6)	49.4 (2.5)	5.8 (0.2)	40.0 (2.9)	5.2 (0.2)	79.4 (6.4)	7.3 (0.3)	60.1 (9.6)	6.4 (0.4
600	100/20	0.6	89.1 (8.6)	7.6 (0.3)	76.8 (15.5)	7.2 (0.5)	49.6 (2.3)	5.9 (0.2)	39.9 (2.9)	5.2 (0.2)	76.2 (7.2)	7.2 (0.3)	57.6 (8.2)	6.3 (0.4)
500	50/35	0.2	87.0 (11.9)	7.6 (0.4)	83.1 (13.5)	7.5 (0.4)	46.9 (2.9)	5.6 (0.2)	44.8 (2.6)	5.5 (0.2)	69.2 (8.8)	6.9 (0.4)	60.5 (9.0)	6.4 (0.4)
500	50/35	0.6	85.3 (11.1)	7.6 (0.3)	82.8 (12.5)	7.4 (0.4)	47.1 (2.9)	5.7 (0.2)	45.1 (2.9)	5.6 (0.2)	67.0 (8.7)	6.9 (0.3)	57.6 (8.3)	6.4 (0.4)
500	50/10	0.2	89.8 (9.0)	7.7 (0.3)	83.5 (13.8)	7.4 (0.4)	50.8 (2.4)	5.9 (0.1)	44.8 (3.0)	5.5 (0.2)	79.8 (6.1)	7.4 (0.2)	60.9 (10.1)	6.5 (0.4)
500	50/10	0.6	89.2 (8.7)	7.7 (0.2)	82.0 (12.2)	7.4 (0.4)	50.8 (2.6)	5.9 (0.2)	44.7 (3.2)	5.6 (0.2)	76.2 (5.7)	7.2 (0.2)	57.9 (8.0)	6.4 (0.4)
350	20/7	0.2	86.0 (13.8)	7.5 (0.4)	80.2 (14.1)	7.4 (0.5)	51.0 (2.4)	5.9 (0.2)	48.4 (2.8)	5.7 (0.2)	75.0 (7.0)	7.2 (0.3)	59.6 (9.1)	6.4 (0.4)
350	20/7	0.6	85.9 (11.4)	7.6 (0.3)	81.6 (13.9)	7.4 (0.4)	50.9 (2.6)	5.9 (0.2)	48.1 (2.6)	5.8 (0.2)	72.0 (6.4)	7.1 (0.3)	57.5 (7.8)	6.4 (0.4)
Nontree-type DTR														
1000	100/20	0.2	94.0 (33.2)	7.8 (1.2)	85.0 (34.4)	7.6 (1.2)	24.3 (2.8)	4.9 (0.2)	18.3 (2.3)	4.4 (0.2)	86.7 (4.6)	7.7 (0.2)	69.8 (11.4)	7.0 (0.4)
1000	100/20	0.6	93.9 (31.0)	7.8 (1.0)	85.9 (32.4)	7.6 (1.2)	25.0 (2.6)	5.0 (0.2)	18.4 (2.0)	4.5 (0.1)	82.0 (5.4)	7.5 (0.2)	61.2 (9.3)	6.7 (0.4)
600	100/20	0.2	71.4 (32.3)	7.3 (1.2)	53.4 (26.4)	6.4 (1.0)	22.0 (3.1)	4.7 (0.2)	15.5 (2.4)	4.2 (0.2)	71.6 (7.4)	7.2 (0.3)	45.4 (10.9)	6.0 (0.5)
600	100/20	0.6	74.9 (30.7)	7.3 (1.2)	55.5 (21.4)	6.4 (1.0)	22.5 (3.3)	4.8 (0.2)	15.6 (2.3)	4.2 (0.2)	66.3 (8.1)	7.0 (0.3)	38.4 (10.5)	5.8 (0.5)
500	50/35	0.2	65.7 (36.0)	7.0 (1.2)	64.3 (34.7)	6.9 (1.2)	19.2 (2.8)	4.5 (0.2)	18.2 (2.7)	4.4 (0.2)	54.4 (9.8)	6.5 (0.4)	43.5 (10.9)	6.0 (0.5)
500	50/35	0.6	72.0 (28.6)	7.2 (1.1)	65.4 (26.0)	7.0 (1.1)	19.6 (2.8)	4.6 (0.2)	18.3 (2.7)	4.5 (0.2)	50.6 (10.3)	6.4 (0.4)	37.7 (10.3)	5.8 (0.5)
500	50/10	0.2	80.9 (31.4)	7.5 (1.2)	64.4 (33.2)	6.9 (1.1)	23.9 (3.4)	4.9 (0.2)	18.1 (2.8)	4.4 (0.2)	72.1 (7.6)	7.2 (0.3)	43.6 (10.5)	6.0 (0.5)
500	50/10	0.6	78.6 (30.3)	7.4 (1.2)	61.5 (27.6)	6.7 (1.1)	24.6 (3.3)	5.0 (0.2)	18.2 (2.7)	4.4 (0.2)	66.3 (7.7)	7.0 (0.3)	37.0 (9.4)	5.8 (0.4)
350	20/7	0.2	72.2 (31.6)	7.3 (1.2)	66.0 (33.6)	7.0 (1.1)	23.9 (4.1)	4.9 (0.3)	20.8 (3.7)	4.6 (0.2)	64.2 (8.1)	6.9 (0.3)	41.2 (9.7)	5.9 (0.4)
350	20/7	0.6	77.2 (28.3)	7.3 (1.1)	68.0 (30.1)	7.1 (1.1)	24.3 (4.2)	5.0 (0.3)	20.7 (3.6)	4.7 (0.2)	60.1 (8.3)	6.8 (0.3)	36.8 (8.7)	5.7 (0.4)

TABLE 4 Characteristics of the analysis cohort. Summary statistics of demographics, treatment, and outcomes for MIMIC-III analysis cohort are included. n = sample size. Stage 1 = 0-3 hours post-admission. Stage 2 = 3-24 hours post-admission. R = restrictive fluid resuscitation (< 30 mL/kg); L = liberal fluid resuscitation (≥ 30 mL/kg); IQR = interquartile range; kg = kilogram; LOS = length of hospital stay; Mech Vent = Mechanical Ventilation; Vasos = Vasopressors; L = liters; mL/kg = milliliters per kilogram; SOFA = sequential organ failure assessment; hrs = hours. Median (IQR) are presented for continuous variables; frequency (percentage) are provided for categorical variables.

	Overall (n=708)	Stage 1		Stage 2	
		Restrictive (n=265)	Liberal (n=443)	Restrictive (n=458)	Liberal (n=250)
Patient Characteristics					
Age (years)	68 [53-81]	70 [55-81]	66 [52-80]	68 [54-81]	68 [53-80]
Gender					
Male	379 (54)	150 (57)	229 (52)	242 (53)	137 (55)
Female	329 (46)	115 (43)	214 (48)	216 (47)	113 (45)
Race/Ethnicity					
White	540 (76)	197 (74)	343 (77)	353 (77)	187 (75)
Nonwhite	168 (24)	68 (26)	100 (23)	105 (23)	63 (25)
Weight (kg)	78 [65-92]	81 [69-98]	75 [63-90]	80 [68-94]	74 [62-89]
BMI (kg/m^2)	27.2 [23.2-31.7]	28.9 [24.2-34.5]	26.8 [22.5-30.7]	28.0 [24.1-32.7]	26.4 [22.0-30.1]
LOS (days)	8.9 [5.7-15.8]	10.5 [6.0-17.5]	8.2 [5.6-14.3]	8.8 [5.5-15.6]	9.7 [6.2-16.9]
0-3 hours post-Admission					
Use of Mech Vent	204 (29)	93 (35)	111 (25)	150 (33)	54 (22)
Use of Vasos	139 (19)	24 (9)	114 (26)	86 (19)	52 (21)
Total Input (L)	3.0 [1.6-5.0]	1.2 [0.8-2.0]	4.0 [3.0-5.5]	2.5 [1.2-4.0]	4.0 [2.5-5.2]
Total Input (mL/kg)	38.0 [20.5-57.8]	16.1 [9.8-22.4]	53.3 [40.0-70.9]	31.7 [15.4-51.1]	49.8 [33.5-70.6]
SOFA (3 hours)	4 [2-6]	4 [2-6]	5 [2-7]	5 [3-6]	4 [2-6]
3-24 hours post-Admission					
Use of Mech Vent	341 (48)	138 (52)	203 (46)	203 (44)	138 (55)
Use of Vasos	294 (42)	80 (30)	214 (48)	148 (32)	146 (58)
Total Input (L)	2.0 [1.0-4.5]	1.4 [0.8-2.5]	3.0 [1.0-5.0]	1.0 [0.5-1.5]	4.5 [3.1-6.0]
Total Input (mL/kg)	16.8 [0.0-44.1]	6.7 [0.0-21.2]	26.8 [6.1-59.0]	5.2 [0.0-15.4]	57.2 [42.2-86.1]
SOFA (24 hours)	5 [3-7]	5 [3-6]	5 [3-8]	5 [3-6]	6 [3-9]



SIM_9155_sepsis_BMI_black_white.png