The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1914–1938
https://doi.org/10.1214/18-AOAS1137

TREE-BASED REINFORCEMENT LEARNING FOR ESTIMATING OPTIMAL DYNAMIC TREATMENT REGIMES[1]

BY YEBIN TAO, LU WANG AND DANIEL ALMIRALL

University of Michigan

Dynamic treatment regimes (DTRs) are sequences of treatment decision rules, in which treatment may be adapted over time in response to the changing course of an individual. Motivated by the substance use disorder (SUD) study, we propose a tree-based reinforcement learning (T-RL) method to directly estimate optimal DTRs in a multi-stage multi-treatment setting. At each stage, T-RL builds an unsupervised decision tree that directly handles the problem of optimization with multiple treatment comparisons, through a purity measure constructed with augmented inverse probability weighted estimators. For the multiple stages, the algorithm is implemented recursively using backward induction. By combining semiparametric regression with flexible tree-based learning, T-RL is robust, efficient and easy to interpret for the identification of optimal DTRs, as shown in the simulation studies. With the proposed method, we identify dynamic SUD treatment regimes for adolescents.

1. Introduction. In many areas of clinical practice, it is often necessary to adapt treatment over time, due to significant heterogeneity in how individuals respond to treatment, as well as to account for the progressive (e.g., cyclical) nature of many chronic diseases and conditions. For example, substance use disorder (SUD) often involves a chronic course of repeated cycles of cessation (or significant reductions in use) followed by relapse [Hser et al. (1997), McLellan et al. (2000)]. However, individuals with SUD are vastly heterogeneous in the course of this disorder, as well as in how they respond to different interventions [Murphy et al. (2007)]. Dynamic treatment regimes (DTRs) [Robins (1986, 1997, 2004),

Received October 2016; revised August 2017.

[1]The development of this article was supported the National Institutes of Health (R01DA039901, P50DA039838 and R01DA015697) and by the Center for Substance Abuse Treatment (CSAT), Substance Abuse and Mental Health Services Administration (SAMHSA) contract 270-07-0191 using data provided by the following grantees: Cannabis Youth Treatment (Study: CYT; CSAT/SAMHSA contracts 270-97-7011, 270-00-6500, 270-2003-00006 and grantees: TI-11317, TI-11321, TI-11323, TI-1324), Adolescent Treatment Model (Study: ATM: CSAT/SAMHSA contracts 270-98-7047, 270-97-7011, 277-00-6500, 270-2003-00006 and grantees: TI-11894, TI-11892, TI-11422, TI-11423, TI-11424, TI-11432), the Strengthening Communities-Youth (Study: SCY; CSAT/SAMHSA contracts 277-00-6500, 270-2003-00006 and grantees: TI-13344, TI-13354, TI-13356), and Targeted Capacity Expansion (Study: TCE; CSAT/SAMHSA contracts 270-2003-00006, 270-2007-00004C, and 277-00-6500 and grantee TI-16400).

Key words and phrases. Multi-stage decision-making, personalized medicine, classification, backward induction, decision tree.

Murphy (2003), Chakraborty and Murphy (2014)] are prespecified sequences of treatment decision rules, designed to help guide clinicians in whether or how, including based on which measures, to adapt (and re-adapt) treatment over time in response to the changing course of an individual. A DTR has multiple stages of treatment, and at each stage, information about a patient's medical history and current disease status can be used to make a treatment recommendation for the next stage. The following is a simple example of a two-stage DTR for adolescents with SUD. First, at treatment program entry, offer adolescents nonresidential (outpatient) treatment for three months, and monitor them for substance use over the course of three months. Second, at the end of three months, if an adolescent has experienced reductions in the frequency of substance use, continue providing outpatient treatment for an additional three months. Otherwise, offer residential (inpatient) treatment for an additional three months. Identification of optimal DTRs offers an effective vehicle for personalized management of diseases, and helps physicians tailor the treatment strategies dynamically and individually based on clinical evidence, thus providing a key foundation for better health care [Wagner et al. (2001)].

Several methods have been developed or modified for the identification of optimal DTRs, which differ in terms of modeling assumptions as well as interpretability, that is, the ease with which it is possible to communicate the decision rules that make up the DTR [Zhang et al. (2015, 2016)]. The interpretability of an estimated optimal DTR is crucial for facilitating applications in medical practice. Commonly used statistical methods include marginal structural models with inverse probability weighting (IPW) [Murphy, van der Laan and Robins (2001), Wang et al. (2012), Hernán, Brumback and Robins (2001)], G-estimation of structural nested mean models [Robins (1994, 1997, 2004)], targeted maximum likelihood estimators [van der Laan and Rubin (2006)] and likelihood-based approaches [Thall et al. (2007)]. To apply these methods, one needs to specify a series of parametric or semiparametric conditional models under a prespecified class of DTRs indexed by unknown parameters, and then search for DTRs that optimize the expected outcome. They often result in estimated optimal DTRs that are highly interpretable. However, in some settings, these methods may be too restrictive; for example, when there is a moderate-to-large number of covariates to consider or when there is no specific class of DTRs of particular interest.

To reduce modeling assumptions, more flexible methods have been proposed. In particular, the problem of developing optimal multi-stage decisions has strong resemblance to reinforcement learning (RL) [Chakraborty and Moodie (2013)]. Unlike supervised learning (SL) (e.g., regression and classification), the desired output value (e.g., the true class or the optimal decision), also known as the *label*, is not observed. The learning agent has to keep interacting with the environment to learn the best decision rule. Such methods include Q-learning [Watkins and Dayan (1992), Sutton and Barto (1998)] and A-learning [Murphy (2003), Schulte et al. (2014)], both of which use backward induction [Bather (2000)] to account

for the delayed (or long-term) effects of earlier-stage treatment decisions. Q- and A-learning rely on maximizing or minimizing an objective function to indirectly infer the optimal DTRs and thus emphasize prediction accuracy of the clinical response model instead of directly optimizing the decision rule [Zhao et al. (2012)]. The modeling flexibility and interpretability of Q- and A-learning depend on the method for optimizing the objective function.

There has also been considerable interest in converting the RL problem to a SL problem so as to utilize existing classification methods. These methods are usually flexible with a nonparametric modeling framework but may introduce additional uncertainty due to the conversion. Their interpretability rests on the choice of the classification approach. For example, Zhao et al. (2015) propose outcome weighted learning (OWL) to transform the optimal DTR problem into an either sequential or simultaneous classification problem, and then apply support vector machines (SVM) [Cortes and Vapnik (1995)]. However, it is difficult to interpret the optimal DTRs estimated by SVM. Moreover, OWL is susceptible to trying to retain the actually observed treatments given a positive outcome, and its estimated individualized treatment rule is affected by a simple shift of the outcome [Zhou et al. (2017)]. For observational data, Tao and Wang (2017) propose a robust method for multi-treatment DTRs, adaptive contrast weighted learning (ACWL), which combines doubly robust augmented IPW (AIPW) estimators with classification algorithms. It avoids the challenging multiple treatment comparisons by utilizing adaptive contrasts that indicate the minimum or maximum expected reduction in the outcome given any sub-optimal treatment. In other words, ACWL ignores information on treatments that lead to neither the minimum or maximum expected reduction in the outcome, likely at the cost of efficiency.

Recently, Laber and Zhao (2015) propose a novel tree-based approach, denoted as LZ hereafter, to directly estimating optimal treatment regimes. Typically, a decision tree is a SL method that uses tree-like graphs or models to map observations about an item to conclusions about the item's target value, for example, the classification and regression tree (CART) algorithm by Breiman et al. (1984). LZ fits the RL task into a decision tree with a purity measure that is unsupervised, and meanwhile maintains the advantages of decision trees, such as simplicity for understanding and interpretation, and capability of handling multiple treatments and various types of outcomes (e.g., continuous or categorical) without distributional assumptions. However, LZ is limited to a single-stage decision problem, and is also susceptible to propensity model misspecification. More recently, Zhang et al. (2015, 2016) and Lakkaraju and Rudin (2017) have applied decision lists to construct interpretable DTRs, which comprise a sequence of "if–then" clauses that map patient covariates to recommended treatments. A decision list can be viewed as a special case of tree-based rules, where the rules are ordered and learned one after another [Rivest (1987)]. These list-based methods are particularly useful when

the goal is not only to gain the maximum health benefits but also to minimize the cost of measuring covariates. However, without cost information, a list-based method may be more restrictive than a tree-based method. On the one hand, to ensure parsimony and interpretability, Zhang et al. (2015, 2016) restrict each rule to involve up to two covariates, which may be problematic for more complex treatment regimes. On the other hand, due to the ordered nature of lists, a later rule is built upon all the previous rules and thus errors can accumulate. In contrast, a decision tree does not require the exploration of a full rule at the very beginning of the algorithm, since the rules are learned at the terminal nodes. Instead of being fully dependent on each other, rules from a decision tree are more related only if they share more parent nodes, which allows more freedom for exploration.

In this paper, we develop a tree-based RL (T-RL) method to directly estimate optimal DTRs in a multi-stage multi-treatment setting, which builds upon the strengths of both ACWL and LZ. First of all, through the use of decision trees, our proposed method is interpretable, capable of handling multinomial or ordinal treatments and flexible for modeling various types of outcomes. Second, thanks to the unique purity measures for a series of unsupervised trees at multiple stages, our method directly incorporates multiple treatment comparisons while maintaining the nature of RL. Last but not least, the proposed method has improved estimation robustness by embedding doubly robust AIPW estimators in the decision tree algorithm.

The remainder of this paper is organized as follows. In Section 2, we formalize the problem of estimating the optimal DTR in a multi-stage multi-treatment setting using the counterfactual framework, derive purity measures for decision trees at multiple stages and describe the recursive tree growing process. The performance of our proposed method in various scenarios is evaluated by simulation studies in Section 3. We further illustrate our method in Section 4 using a case study to identify optimal dynamic substance abuse treatment regimes for adolescents. Finally, we conclude with some discussions and suggestions for future research in Section 5.

2. Tree-based reinforcement learning (T-RL).

2.1. *Dynamic treatment regimes* (*DTRs*). Consider a multi-stage decision problem with T decision stages and K_j ($K_j \geq 2$) treatment options at the jth ($j = 1, \ldots, T$) stage. Data could come from either a randomized trial or an observational study. Let A_j denote the multi-categorical treatment indicator with observed value $a_j \in \mathcal{A}_j = \{1, \ldots, K_j\}$. In the SUD data, treatment is multi-categorical with options being residential, non-residential or no treatment. Let $\mathbf{X}_j$ denote the vector of patient characteristics history just prior to treatment assignment A_j, and $\mathbf{X}_{T+1}$ denote the entire characteristics history up to the end of stage T. Let R_j be the reward (e.g., reduction in the frequency of substance use)

following A_j, which could depend on the covariate history $\mathbf{X}_j$ and treatment history $A_1, \ldots, A_j$, and is also a part of the covariate history $\mathbf{X}_{j+1}$. We consider the overall outcome of interest as $Y = f(R_1, \ldots, R_T)$, where $f(\cdot)$ is a prespecified function (e.g., sum), and we assume that Y is bounded; higher values of Y are preferable. The observed data are $\{(A_{1i}, \ldots, A_{Ti}, \mathbf{X}_{T+1,i}^\top)\}_{i=1}^n$, assumed to be independent and identically distributed for n subjects from a population of interest. For brevity, we suppress the subject index i in the following text when no confusion exists.

A DTR is a sequence of individualized treatment rules, $\mathbf{g} = (g_1, \ldots, g_T)$, where g_j is a mapping from the domain of covariate and treatment history $\mathbf{H}_j = (A_1, \ldots, A_{j-1}, \mathbf{X}_j^\top)^\top$ to the domain of treatment assignment A_j, and we set $A_0 = \varnothing$. To define and identify the optimal DTR, we consider the counterfactual framework for causal inference [Robins (1986)].

At stage T, let $Y^*(A_1, \ldots, A_{T-1}, a_T)$, or $Y^*(a_T)$ for brevity, denote the counterfactual outcome for a patient treated with $a_T \in \mathcal{A}_T$ conditional on previous treatments $(A_1, \ldots, A_{T-1})$, and define $Y^*(g_T)$ as the counterfactual outcome under regime g_T, that is,

$$Y^*(g_T) = \sum_{a_T=1}^{K_T} Y^*(a_T) I\{g_T(\mathbf{H}_T) = a_T\}.$$

The performance of g_T is measured by the counterfactual mean outcome $E\{Y^*(g_T)\}$, and the optimal regime, g_T^{opt}, satisfies $E\{Y^*(g_T^{\text{opt}})\} \geq E\{Y^*(g_T)\}$ for all $g_T \in \mathcal{G}_T$, where $\mathcal{G}_T$ is the class of all potential regimes. To connect the counterfactual outcomes with the observed data, we make the following three standard assumptions [Murphy, van der Laan and Robins (2001), Robins and Hernán (2009), Orellana, Rotnitzky and Robins (2010)].

ASSUMPTION 1 (Consistency). The observed outcome is the same as the counterfactual outcome under the treatment a patient is actually given, that is, $Y = \sum_{a_T=1}^{K_T} Y^*(a_T) I(A_T = a_T)$, where $I(\cdot)$ is the indicator function that takes the value 1 if $\cdot$ is true and 0 otherwise. It also implies that there is no interference between subjects.

ASSUMPTION 2 (No unmeasured confounding). Treatment A_T is randomly assigned with probability possibly dependent on $\mathbf{H}_T$, that is,

$$\{Y^*(1), \ldots, Y^*(K_T)\} \perp\!\!\!\perp A_T \mid \mathbf{H}_T,$$

where $\perp\!\!\!\perp$ denotes statistical independence.

ASSUMPTION 3 (Positivity). There exist constants $0 < c_0 < c_1 < 1$ such that, with probability 1, the propensity score $\pi_{a_T}(\mathbf{H}_T) = \Pr(A_T = a_T | \mathbf{H}_T) \in (c_0, c_1)$.

Following the derivation in Tao and Wang (2017) under the foregoing three assumptions, we have

$$E\{Y_T^*(g_T)\} = E_{\mathbf{H}_T}\left[\sum_{a_T=1}^{K_T} E(Y|A_T = a_T, \mathbf{H}_T) I\{g_T(\mathbf{H}_T) = a_T\}\right],$$

where $E_{\mathbf{H}_T}(\cdot)$ denotes expectation with respect to the marginal joint distribution of the observed data $\mathbf{H}_T$. If we denote the conditional mean $E(Y|A_T = a_T, \mathbf{H}_T)$ as $\mu_{T,a_T}(\mathbf{H}_T)$, we have

$$g_T^{\text{opt}} = \arg\max_{g_T \in \mathcal{G}_T} E_{\mathbf{H}_T}\left[\sum_{a_T=1}^{K_T} \mu_{T,a_T}(\mathbf{H}_T) I\{g_T(\mathbf{H}_T) = a_T\}\right]. \tag{2.1}$$

At stage j, $T-1 \geq j \geq 1$, g_j^{opt} can be expressed in terms of the observed data via backward induction [Bather (2000)]. Following Murphy (2005) and Moodie, Chakraborty and Kramer (2012), we define a stage-specific pseudo-outcome PO_j for estimating g_j^{opt}, which is a predicted counterfactual outcome under optimal treatments at all future stages, also known as the value function. Specifically, we have

$$PO_j = E\{Y^*(A_1, \ldots, A_j, g_{j+1}^{\text{opt}}, \ldots, g_T^{\text{opt}})\},$$

or in a recursive form,

$$PO_j = E\{PO_{j+1}|A_{j+1} = g_{j+1}^{\text{opt}}(\mathbf{H}_{j+1}), \mathbf{H}_{j+1}\}$$

and we set $PO_T = Y$.

For $a_j = 1, \ldots, K_j$, let $\mu_{j,a_j}(\mathbf{H}_j)$ denote the conditional mean $E[PO_j|A_j = a_j, \mathbf{H}_j]$, and we have $PO_j = \mu_{j+1,g_{j+1}^{\text{opt}}}(\mathbf{H}_{j+1})$. Let $PO_j^*(a_j)$ denote the counterfactual pseudo-outcome for a patient with treatment a_j at stage j. For the three assumptions, we have *positivity* as $PO_j = \sum_{a_j=1}^{K_j} PO_j^*(a_j) I(A_j = a_j)$, *no unmeasured confounding* as $\{PO_j^*(1), \ldots, PO_j^*(K_j)\} \perp\!\!\!\perp \mathbf{H}_j$ and *positivity* as $\pi_{a_j}(\mathbf{H}_j) = \Pr(A_j = a_j|\mathbf{H}_j)$ bounded away from zero and one. Under these three assumptions, the optimization problem at stage j, among all potential regimes $\mathcal{G}_j$, can be written as

$$g_j^{\text{opt}} = \arg\max_{g_j \in \mathcal{G}_j} E_{\mathbf{H}_j}\left[\sum_{a_j=1}^{K_j} \mu_{j,a_j}(\mathbf{H}_j) I\{g_j(\mathbf{H}_j) = a_j\}\right]. \tag{2.2}$$

2.2. *Purity measures for decision trees at multiple stages.* We propose to use a tree-based method to solve (2.1) and (2.2). Typically, a CART tree is a binary decision tree constructed by splitting a parent node into two child nodes repeatedly, starting with the root node which contains the entire learning samples. The

basic idea of tree growing is to choose a split among all possible splits at each node so that the resulting child nodes are the purest (e.g., having the lowest misclassification rate). Thus the purity or impurity measure is crucial to the tree growing. Traditional classification and regression trees are SL methods, with the goal of inferring a function that describes the relationship between the outcome and covariates. The desired output value, also known as the *label*, is observed and can be used directly to measure purity. Commonly used impurity measures include Gini index and information index for categorical outcomes, and least squares deviation for continuous outcomes [Breiman et al. (1984)].

However, the estimation target of a DTR problem, which is the optimal treatment for a patient with characteristics $\mathbf{H}_j$ at stage j, that is, $g_j^{\text{opt}}(\mathbf{H}_j)$, $j = 1, \ldots, T$, is not directly observed. Information about $g_j^{\text{opt}}(\mathbf{H}_j)$ can only be inferred indirectly through the observed treatments and outcomes. Using the causal framework and the foregoing three assumptions, we can pool over all subject-level data to estimate the counterfactual mean outcomes given all possible treatments. With the overall goal of maximizing the counterfactual mean outcome in the entire population of interest, the selected split at each node should also improve the counterfactual mean outcome, which can serve as a measure of purity in DTR trees. Figure 1 illustrates a decision tree for a single-stage ($T = 1$) optimal treatment rule with $\mathcal{A} = \{0, 1, 2\}$. Let $\Omega_m, m = 1, 2, \ldots,$ denote the nodes which are regions defined by the covariate space following all precedent binary splits, with the root node $\Omega_1 = \mathbb{R}^p$ (p is the covariate dimension). We number the rectangular region $\Omega_m, m \geq 2$, so that its parent node is $\Omega_{\lceil m/2 \rceil}$, where $\lceil \cdot \rceil$ means taking the smallest integer not less than $\cdot$. Figure 1 shows the chosen covariate and best split at each node, as well as the counterfactual mean outcome after assigning a single optimal treatment to that node. The splits are selected to increase the counterfactual mean outcome. At the root node, if we select a single treatment for all subjects, treatment

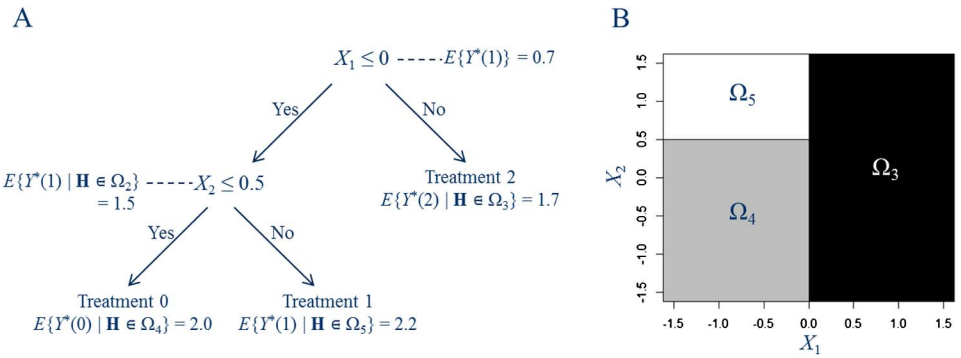

FIG. 1. (A) *A decision tree for optimal treatment rules and the expected counterfactual outcome by assigning a single best treatment to each node that represents a subset covariate space.* (B) *Regions divided by the terminal nodes in the decision tree indicating different optimal treatments.*

1 is the most beneficial overall, yielding a counterfactual mean outcome of 0.7. Splitting via X_1 and X_2, the optimal regime g^{opt} is to assign treatment 2 to region $\Omega_3 = \{X_1 > 0\}$, treatment 0 to region $\Omega_4 = \{X_1 \le 0, X_2 \le 0.5\}$, and treatment 1 to region $\Omega_5 = \{X_1 \le 0, X_2 > 0.5\}$. We can see that this tree is fundamentally different from a CART tree as it does not attempt to describe the relationship between the outcome and covariates or the rule for the assignment of the observed treatments, and instead it describes the rule by which treatments should be assigned to future subjects in order to maximize the purity, which is the counterfactual mean outcome.

Laber and Zhao (2015) propose a measure of node purity based on the IPW estimator of the counterfactual mean outcome [Zhang et al. (2012), Zhao et al. (2012)],

$$E\{Y^*(g)\} = E_{\mathbf{H}}\left[\frac{I(A = g(\mathbf{H}))}{\pi_A(\mathbf{H})}Y\right],$$

for a single-stage ($T = 1$, omitted for brevity) decision problem. Given known propensity score $\pi_A(\mathbf{H})$, they propose a purity measure $\mathcal{P}^{\text{LZ}}(\Omega, \omega)$ as

$$\max_{a_1,a_2\in\mathcal{A}} \frac{\mathbb{P}_n[\frac{\{Y-\hat{m}(\mathbf{H})\}I\{A=g_{\omega,a_1,a_2}(\mathbf{H})\}}{\pi_A(\mathbf{H})} I(\mathbf{H}\in\Omega)]}{\mathbb{P}_n[\frac{I\{A=g_{\omega,a_1,a_2}(\mathbf{H})\}}{\pi_A(\mathbf{H})} I(\mathbf{H}\in\Omega)]},$$

where $\mathbb{P}_n$ is the empirical expectation operator, $\hat{m}(\mathbf{H})$ is $\max_{a\in\mathcal{A}}\hat{\mu}_a(\mathbf{H})$ with $\mu_a(\mathbf{H}) = E(Y|A = a, \mathbf{H})$, Ω denotes the node to be split, ω and ω^c is a partition of Ω, and for a given partition ω and ω^c, g_{ω,a_1,a_2} denotes the decision rule that assigns treatment a_1 to subjects in ω and treatment a_2 to subjects in ω^c. $\mathcal{P}^{\text{LZ}}(\Omega, \omega)$ is the estimated counterfactual mean outcome for node Ω by the best decision rule that assigns a single treatment to all subjects in ω and a second treatment to all subjects in ω^c. However, in an observational study where $\pi_A(\mathbf{H})$ has to be estimated, $\mathcal{P}^{\text{LZ}}(\Omega, \omega)$ is subject to misspecification of the propensity model. Moreover, as the node size decreases, the IPW-based purity measure will become less stable.

To improve robustness, we propose to use an AIPW estimator for the counterfactual mean outcome as in Tao and Wang (2017). By regarding the K treatment options as K arbitrary missing data patterns [Rotnitzky, Robins and Scharfstein (1998)], the AIPW estimator for $E\{Y^*(a)\}$ is $\mathbb{P}_n\{\hat{\mu}_a^{\text{AIPW}}(\mathbf{H})\}$, with

$$\hat{\mu}_a^{\text{AIPW}}(\mathbf{H}) = \frac{I(A = a)}{\hat{\pi}_a(\mathbf{H})}Y + \left\{1 - \frac{I(A = a)}{\hat{\pi}_a(\mathbf{H})}\right\}\hat{\mu}_a(\mathbf{H}). \tag{2.3}$$

Under the foregoing three assumptions, $\mathbb{P}_n\{\hat{\mu}_a^{\text{AIPW}}(\mathbf{H})\}$ is a consistent estimator of $E\{Y^*(a)\}$ if either the propensity model $\pi_a(\mathbf{H})$ or the conditional mean model $\mu_a(\mathbf{H})$ is correctly specified, and thus the method is doubly robust.

In our multi-stage setting, for stage T, given estimated conditional mean $\hat{\mu}^{\mathrm{AIPW}}_{T,a_T}(\mathbf{H}_T)$ and estimated propensity score $\hat{\pi}_{T,A_T}(\mathbf{H}_T)$, the proposed estimator of $E\{Y^*_T(g_T)\}$ is

$$\mathbb{P}_n\Bigg[\sum_{a_T=1}^{K_T}\hat{\mu}^{\mathrm{AIPW}}_{T,a_T}(\mathbf{H}_T)I\{g_T(\mathbf{H}_T)=a_T\}\Bigg] = \mathbb{P}_n\Bigg[\frac{I(A_T=g_T(\mathbf{H}_T))}{\hat{\pi}_{T,A_T}(\mathbf{H}_T)}Y+\Bigg\{1-\frac{I(A_T=g_T(\mathbf{H}_T))}{\hat{\pi}_{T,A_T}(\mathbf{H}_T)}\Bigg\}\hat{\mu}_{T,g_T}(\mathbf{H}_T)\Bigg],$$

which has the augmented term in addition to the IPW estimator used by Laber and Zhao (2015). Similarly, for stage j $(T-1\le j\le 1)$, the proposed estimator of $E\{PO^*_j(g_j)\}$ is

$$\mathbb{P}_n\Bigg[\frac{I(A_j=g_j(\mathbf{H}_j))}{\hat{\pi}_{j,A_j}(\mathbf{H}_j)}\widehat{PO}_j+\Bigg\{1-\frac{I(A_j=g_j(\mathbf{H}_j))}{\hat{\pi}_{j,A_j}(\mathbf{H}_j)}\Bigg\}\hat{\mu}_{j,g_j}(\mathbf{H}_j)\Bigg],$$

where $\hat{\pi}_{j,A_j}(\mathbf{H}_j)$ is the estimated propensity score, $\hat{\mu}_{j,a_j}(\mathbf{H}_j)$ is the estimated conditional mean and $\widehat{PO}_j=\hat{\mu}_{j+1,\hat{g}^{\mathrm{opt}}_{j+1}}(\mathbf{H}_{j+1})$ is the estimated pseudo-outcome.

Our proposed method maximizes the counterfactual mean outcome through each of the nodes. For a given partition ω and ω^c of node Ω, let g_{j,ω,a_1,a_2} denote the decision rule that assigns treatment a_1 to subjects in ω and treatment a_2 to subjects in ω^c at stage j $(T\le j\le 1)$, and we define the purity measure $\mathcal{P}_j(\Omega,\omega)$ as

$$\max_{a_1,a_2\in\mathcal{A}_j}\mathbb{P}_n\Bigg[\sum_{a_j=1}^{K_j}\hat{\mu}^{\mathrm{AIPW}}_{j,a_j}(\mathbf{H}_j)I\{g_{j,\omega,a_1,a_2}(\mathbf{H}_j)=a_j\}I(\mathbf{H}_j\in\Omega)\Bigg].$$

We can see that $\mathcal{P}_j(\Omega,\omega)$ is the estimated counterfactual mean outcome for node Ω and it works as the performance measure for the best decision rule which assigns a single treatment to each of the two arms under the partition ω. Comparing $\mathcal{P}_j(\Omega,\omega)$ and $\mathcal{P}^{\mathrm{LZ}}(\Omega,\omega)$, the primary difference is in the underlying estimator for the counterfactual mean outcome. Another difference is that in $\mathcal{P}_j(\Omega,\omega)$, one is utilizing all subjects at node Ω with the counterfactual outcomes $\hat{\mu}^{\mathrm{AIPW}}_{j,a_j}(\mathbf{H}_j)$ calculated using all samples at the root node, while in $\mathcal{P}^{\mathrm{LZ}}(\Omega,\omega)$, one is only using a subset of subjects, depending on compatibility to $g_{\omega,a_1,a_2}(\mathbf{H}_j)$, which is why there is a denominator in $\mathcal{P}^{\mathrm{LZ}}(\Omega,\omega)$ but not in $\mathcal{P}_j(\Omega,\omega)$. These differences may lead to better stability for $\mathcal{P}_j(\Omega,\omega)$.

2.3. *Recursive partitioning*. As we have mentioned, the purity measures for our T-RL are different from the ones in supervised decision trees. However, after defining $\mathcal{P}_j(\Omega,\omega)$, $j=1,\dots,T$, the recursive partitioning to grow the tree is similar. Each split depends on the value of only one covariate. A nominal covariate

with C categories has $2^{C-1}-1$ possible splits and an ordinal or continuous covariate with L different values has $L-1$ unique splits. Therefore, at a given node Ω, a possible split ω indicates either a subset of categories for a nominal covariate or values no larger than a threshold for an ordinal or continuous covariate. The best split ω^{opt} is chosen to maximize the improvement in the purity, $\mathcal{P}_j(\Omega,\omega)-\mathcal{P}_j(\Omega)$, where $\mathcal{P}_j(\Omega)$ means to assign a single best treatment to all subjects in Ω without splitting. It is straightforward to see that $\mathcal{P}_j(\Omega,\omega)\geq\mathcal{P}_j(\Omega)$. In order to control overfitting as well as to make meaningful splitting, a positive constant λ is given to represent a threshold for practical significance and another positive integer n_0 is given as the minimal node size which is dictated by problem-specific considerations. Under these conditions, we first evaluate the following three *Stopping Rules* for node Ω.

RULE 1. If the node size is less than $2n_0$, the node will not be split.

RULE 2. If all possible splits of a node result in a child node with size smaller than n_0, the node will not be split.

RULE 3. If the current tree depth reaches the user-specified maximum depth, the tree growing process will stop.

If none of the foregoing *Stopping Rules* are met, we compute the best split by

$$\hat{\omega}^{\text{opt}}=\arg\max_{\omega}\bigl[\mathcal{P}_j(\Omega,\omega):\min\bigl\{n\mathbb{P}_n I(\mathbf{H}_j\in\omega),n\mathbb{P}_n I(\mathbf{H}_j\in\omega^c)\bigr\}\geq n_0\bigr].$$

Before deciding whether or not to split Ω into ω and ω^c, we evaluate the following *Stopping Rule* 4.

RULE 4. If the maximum purity improvement $\mathcal{P}_j(\Omega,\hat{\omega}^{\text{opt}})-\mathcal{P}_j(\Omega)$ is less than λ, the node will not be split.

We split Ω into ω and ω^c if none of the four stopping rules apply.

When there is no clear scientific guidance on λ to indicate practical significance, one approach is to choose a relatively small positive value to build a complete tree and then prune the tree back in order to minimize a measure of cost for the tree. Following the CART algorithm, the cost is a measure of the total impurity of the tree with a penalty term on the number of terminal nodes, and the complexity parameter for the penalty term can be tuned by cross-validation (CV) [Breiman et al. (1984)]. Alternatively, we propose to select λ directly by CV, similar to the method by Laber and Zhao (2015). As a direct measure of purity is not available in RL, we again incorporate the idea of maximizing the counterfactual mean outcome and use a 10-fold CV estimator of the counterfactual mean outcome. Theoretically, CV can be conducted at each stage separately and one can use a potentially different

λ for each stage. To reduce modeling uncertainty in the pseudo-outcomes and also simplify the process, we carry out CV only at stage T using the overall outcome Y directly. Specifically, we use nine subsamples as training data to estimate the function of $\mu_{T,a_T}(\cdot)$ following (2.3) and $g_T^{\text{opt}}(\cdot)$ using T-RL for a given λ, and then plug in $\mathbf{H}_T$ of the remaining subsample to get $\hat{\mu}_{T,a_T}^{\text{AIPW,CV}}(\mathbf{H}_T)$ and $\hat{g}_T^{\text{opt,CV},\lambda}(\mathbf{H}_T)$. We repeat the process 10 times with each subsample being the test data once. Then the CV-based counterfactual mean outcome under λ is

$$\hat{E}\{Y^*(\hat{g}_T^{\text{opt,CV},\lambda})\} = \mathbb{P}_n\left[\sum_{a_T=1}^{K_T} \hat{\mu}_{T,a_T}^{\text{AIPW,CV}}(\mathbf{H}_T) I\{\hat{g}_T^{\text{opt,CV},\lambda}(\mathbf{H}_T) = a_T\}\right],$$

and the best value for λ is $\hat{\lambda} = \arg\max_\lambda \hat{E}\{Y^*(\hat{g}_T^{\text{opt,CV},\lambda})\}$. As the scale of the outcome affects the scale of $\mathcal{P}_j(\Omega, \omega) - \mathcal{P}_j(\Omega)$, we search over a sequence of candidate λ's as a sequence of percentages of $\mathcal{P}_T(\Omega_1)$, that is, the estimated counterfactual mean outcome under a single best treatment for all subjects (Ω_1 is the root node).

2.4. *Implementation of T-RL.* The AIPW estimator $\hat{\mu}_{j,a_j}^{\text{AIPW}}(\mathbf{H}_j)$, $j = 1, \ldots, T$, $a_j = 1, \ldots, K_j$, consists of three parts to be estimated, the pesudo-outcome PO_j, the propensity score $\pi_{j,a_j}(\mathbf{H}_j)$ and the conditional mean model $\mu_{j,a_j}(\mathbf{H}_j)$.

We start the estimation with stage T and conduct backward induction. At stage T, we use the outcome Y directly, that is, $PO_T = Y$. For stage $j, T-1 \geq j \geq 1$, given a cumulative outcome (e.g., the sum of longitudinally observed values or a single continuous final outcome), we use a modified version of pseudo-outcomes to reduce accumulated bias from the conditional mean models [Huang et al. (2015)]. Instead of using only the model-based values under optimal future treatments, that is, $\mu_{j+1,g_{j+1}^{\text{opt}}}(\mathbf{H}_{j+1})$, we use the actual observed outcomes plus the expected future loss due to sub-optimal treatments, which means

$$PO'_j = PO'_{j+1} + \mu_{j+1,g_{j+1}^{\text{opt}}}(\mathbf{H}_{j+1}) - \mu_{j+1,A_{j+1}}(\mathbf{H}_{j+1}),$$

where $\mu_{j+1,g_{j+1}^{\text{opt}}}(\mathbf{H}_{j+1}) - \mu_{j+1,a_{j+1}}(\mathbf{H}_{j+1})$ is the expected loss due to sub-optimal treatments at stage $j+1$ for a given patient, which is zero if $g_{j+1}^{\text{opt}}(\mathbf{H}_{j+1}) = A_{j+1}$ and positive otherwise. Given $PO'_T = Y$, it is easy to see that

$$PO'_j = Y + \sum_{t=j+1}^{T} \{\mu_{t,g_t^{\text{opt}}}(\mathbf{H}_t) - \mu_{t,A_t}(\mathbf{H}_t)\}.$$

This modification leads to more robustness against model misspecification and is less likely to accumulate bias from stage to stage during backward induction [Huang et al. (2015)]. In our simulations, we estimate PO'_j by using random forests-based conditional mean estimates [Breiman (2001)].

The propensity score $\pi_{j,a_j}(\mathbf{H}_j)$ can be estimated via multinomial logistic regression [Menard (2002)]. A working model could include linear main effect terms for all variables in $\mathbf{H}_j$. Summary variables or interaction terms may also be included based on scientific knowledge.

The conditional mean estimate $\hat{\mu}_{j,a_j}(\mathbf{H}_j)$ in the augmentation term of $\hat{\mu}^{\mathrm{AIPW}}_{j,a_j}(\mathbf{H}_j)$ can be obtained from a parametric regression model. For continuous outcomes, a simple and oftentimes reasonable example is the parametric linear model with coefficients dependent on treatment:

$$E(\widehat{PO}'_j | A_j, \mathbf{H}_j) = \sum_{a_j=1}^{K_j} (\beta_{a_j}^{\top} \mathbf{H}_j) I(A_j = a_j), \tag{2.4}$$

where β_a is a parameter vector for $\mathbf{H}_j$ under treatment a_j. For binary and count outcomes, one may extend the method by using generalized linear models. For survival outcomes with noninformative censoring, it is possible to use an accelerated failure time model to predict survival time for all patients. Survival outcomes with more complex censoring issues are beyond the scope of the current study.

The T-RL algorithm starting with stage $j = T$ is carried out as follows:

STEP 1. Obtain AIPW estimates $\hat{\mu}^{\mathrm{AIPW}}_{j,a_j}(\mathbf{H}_j), a_j = 1, \ldots, K_j$, using full data.

STEP 2. At root node $\Omega_{j,m}, m = 1$, set values for λ and n_0.

STEP 3. At node $\Omega_{j,m}$, evaluate the four *Stopping Rules*. If any of the *Stopping Rules* is satisfied, assign a single best treatment

$$\arg\max_{a_j \in \mathcal{A}_j} \mathbb{P}_n\big[\hat{\mu}^{\mathrm{AIPW}}_{j,a_j}(\mathbf{H}_j) I(\mathbf{H}_j \in \Omega_{j,m})\big]$$

to all subject in $\Omega_{j,m}$. Otherwise, split $\Omega_{j,m}$ into child nodes $\Omega_{j,2m}$ and $\Omega_{j,2m+1}$ by $\hat{\omega}^{\mathrm{opt}}$.

STEP 4. Set $m = m + 1$ and repeat Step 3 until all nodes are terminal.

STEP 5. If $j > 1$, set $j = j - 1$ and repeat steps 1 to 4. If $j = 1$, stop.

Similar to the CART algorithm, T-RL is greedy as it chooses splits only at the current node for purity improvement, which may not lead to a global maximum. One way to potentially enhance the performance is lookahead [Murthy and Salzberg (1995)]. We test this in our simulation by fixed-depth lookahead: evaluating the purity improvement after splitting the parent node as well as its two child nodes, comparing the total purity improvement after splitting up to four nodes to the purity improvement without splitting the parent node, and finally deciding whether or not to split the parent node. We denote this method as T-RL-LH.

3. Simulation studies. We conduct simulation studies to investigate the performance of our proposed method. We set all regression models μ to be misspecified, which is the case for most real data applications, while allowing the specification of the propensity model π be either correct (e.g., randomized trials) or incorrect (e.g., most observational studies). We consider first a single-stage scenario so as to facilitate the comparison with existing methods, particularly Laber and Zhao (2015), and then a multi-stage scenario. For each scenario, we consider sample sizes of either 500 or 1000 for the training datasets and 1000 for the test datasets, and repeat the simulation 500 times. We use the training datasets to estimate the optimal regime and then predict the optimal treatments in the test datasets, where the underlying truth is known. We denote the percentage of subjects correctly classified to their optimal treatments as opt%. We also use the true outcome model and the estimated optimal regime in the test datasets to estimate the counterfactual mean outcome, denoted as $\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$. For both scenarios, we generate five baseline covariates $X_1, \ldots, X_5$ according to $N(0, 1)$, and for Scenario 1, we further consider a setting with additional covariates $X_6, \ldots, X_{20}$ simulated independently from $N(0, 1)$.

3.1. *Scenario* 1: $T = 1$ *and* $K = 3$. In Scenario 1, we consider a single stage with three treatment options and sample size of 500. The treatment A is set to take values in $\{0, 1, 2\}$, and we generate it from Multinomial(π_0, π_1, π_2), with $\pi_0 = 1/\{1 + \exp(0.5X_1 + 0.5X_4) + \exp(-0.5X_1 + 0.5X_5)\}$, $\pi_1 = \exp(0.5X_1 + 0.5X_4)/\{1 + \exp(0.5X_1 + 0.5X_4) + \exp(-0.5X_1 + 0.5X_5)\}$ and $\pi_2 = 1 - \pi_0 - \pi_1$. The underlying optimal regime is

$$g^{\text{opt}}(\mathbf{H}) = \begin{cases} 0 & X_1 \le 0, X_2 \le 0.5, \\ 2 & X_1 > 0, X_3 \le 0.5, \\ 1 & \text{otherwise.} \end{cases}$$

For the outcomes, we first consider equal penalties for sub-optimal treatments through outcome generating model (a), which is

$$Y = 1 + X_4 + X_5 + \sum_{a=0}^{2} \big[I(A = a)\{2I(g^{\text{opt}} = a) - 1\}\big] + \varepsilon.$$

Then we consider varying penalties for sub-optimal treatments through outcome generating model (b), which is

$$\begin{aligned} Y = 0.79 + X_4 + X_5 + 2I(A = 0)\{2I(g^{\text{opt}} = 0) - 1\} \\ + 1.5I(A = 2)\{2I(g^{\text{opt}} = 2) - 1\} + \varepsilon. \end{aligned}$$

In both outcome models, we have $\varepsilon \sim N(0, 1)$ and $E\{Y^*(g^{\text{opt}})\} = 2$.

In the application of the proposed T-RL algorithm, we consider both a correctly specified model $\log(\pi_d/\pi_0) = \beta_{0d} + \beta_{1d}X_1 + \beta_{2d}X_4 + \beta_{3d}X_5, d = 1, 2$, and an

incorrectly specified one $\log(\pi_d/\pi_0) = \beta_{0d} + \beta_{1d}X_2 + \beta_{2d}X_3$. We also apply T-RL-LH to Scenario 1 as mentioned in Section 2.4. For comparison, we use both the linear regression-based and random forests-based conditional mean models to infer the optimal regimes, which we denote as RG and RF, respectively. We also apply the tree-based method LZ by Laber and Zhao (2015). Furthermore, we apply the OWL method by Zhao et al. (2012), and the ACWL algorithm by Tao and Wang (2017), denoted as ACWL-C_1 and ACWL-C_2, where C_1 and C_2 indicate respectively the minimum and maximum expected loss in the outcome given any sub-optimal treatment for each patient. Given outcome model (a), all sub-optimal treatments have the same expected loss in the outcome and we expect ACWL to perform similarly well as T-RL. However, given outcome model (b) when the sub-optimal treatments have different expected losses in the outcome, we expect T-RL to perform better as it incorporates multiple treatment comparison. Both OWL and ACWL are implemented using the R package *rpart* for classification.

Table 1 summarizes the performances of all methods considered in Scenario 1 with five baseline covariates. We present the percentage of subjects correctly classified to their optimal treatments in the testing datasets, denoted as opt%, and the expected counterfactual outcome obtained using the true outcome model and the estimated optimal regime, denoted as $\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$. opt% shows on average how accurately the estimated optimal regime assigns future patients to their true optimal treatments and $\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$ shows how much the entire population of interest will benefit from following $\hat{g}^{\text{opt}}$. T-RL-LH has the best performance among all the methods considered, classifying over 93% of subjects to their optimal treatments. However, lookahead has led to significant increase in computational time compared to T-RL, while the improvement is only moderate with $\leq 1\%$ more subjects being correctly classified. T-RL also has an estimated counterfactual mean outcome very close to the true value 2. As expected, ACWL-C_1 and ACWL-C_2 have performances comparable to T-RL under outcome model (a) with equal penalties for treatment misclassification, and the performance discrepancy gets larger under outcome model (b) with varying penalties, due to the approximation by adaptive contrasts C_1 and C_2. Similar results can be found in the Supplementary Table S1. LZ, using an IPW-based decision tree, works well only when the propensity score model is correctly specified and is less efficient than T-RL with larger empirical standard deviations (SDs). In contrast, T-RL-LH, T-RL, ACWL-C_1 and ACWL-C_2 are all highly robust to model misclassification, thanks to the combination of doubly robust AIPW estimators and flexible machine learning methods. OWL performs far worse than all other competing methods likely due to the low percentage of truly optimal treatments in the observed treatments, the shift in the outcome, which was intended to ensure positive weights, and its moderate efficiency.

After the inclusion of more noise covariates in Table 2, all methods have worse performances compared to Table 1, with RF suffering the most. T-RL and T-RL-LH have the slightest decreases in opt% and $\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$, showing satisfactory stability against noise interference. Thanks to the built-in variable selection feature

TABLE 1

Simulation results for Scenario 1 *with a single stage, three treatment options and five baseline covariates* (500 *replications*, $n = 500$). π *is the propensity score model*. (a) *and* (b) *indicate equal and varying penalties for treatment misclassification in the generative outcome model*. opt% *shows the empirical mean and standard deviation* (*SD*) *of the percentage of subjects correctly classified to their optimal treatments*. $\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$ *shows the empirical mean and SD of the expected counterfactual outcome obtained using the true outcome model and the estimated optimal regime*. $E\{Y^*(g^{\text{opt}})\} = 2$

π	Method	(a)		(b)	
		opt%	$\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$	opt%	$\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$
–	RG	74.2 (2.3)	1.49 (0.07)	68.8 (4.0)	1.42 (0.09)
	RF	75.3 (4.5)	1.51 (0.11)	81.1 (4.5)	1.69 (0.10)
Correct	OWL	44.3 (7.6)	0.89 (0.16)	47.1 (8.1)	0.89 (0.21)
	LZ	91.5 (7.5)	1.83 (0.16)	89.4 (9.5)	1.81 (0.18)
	ACWL-C_1	93.7 (4.1)	1.87 (0.10)	89.1 (5.3)	1.80 (0.11)
	ACWL-C_2	94.7 (3.3)	1.89 (0.09)	87.8 (5.5)	1.79 (0.11)
	T-RL	97.2 (3.3)	1.95 (0.08)	95.1 (5.6)	1.92 (0.11)
	T-RL-LH	97.5 (3.1)	1.96 (0.08)	96.1 (4.0)	1.94 (0.08)
Incorrect	OWL	33.5 (6.0)	0.67 (0.13)	36.7 (5.7)	0.64 (0.19)
	LZ	87.8 (12.0)	1.75 (0.25)	81.8 (14.7)	1.68 (0.27)
	ACWL-C_1	92.1 (4.7)	1.84 (0.10)	87.9 (5.6)	1.79 (0.11)
	ACWL-C_2	94.7 (3.4)	1.89 (0.09)	86.5 (6.1)	1.78 (0.12)
	T-RL	97.8 (1.8)	1.94 (0.06)	92.9 (7.2)	1.89 (0.13)
	T-RL-LH	98.2 (1.6)	1.95 (0.06)	93.7 (6.2)	1.91 (0.10)

RG, linear regression; RF, random forests; OWL, outcome weighted learning; LZ, method by Laber and Zhao (2015); ACWL-C_1 and ACWL-C_2, method by Tao and Wang (2017); T-RL, tree-based reinforcement learning; T-RL-LH, T-RL with one step lookahead.

of decision trees, LZ and ACWL with CART are also relatively stable. Figure 2 shows the density plots for $\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$ under outcome model (b), with each panel showing correctly or incorrectly specified propensity model and five or 20 baseline covariates. LZ is the least efficient method with the density plots more spread out. T-RL has the least density in lower values of $\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$ and the highest density in higher values.

3.2. *Scenario* 2: $T = 2$ *and* $K_1 = K_2 = 3$. In Scenario 2, we generate data under a two-stage DTR with three treatment options at each stage and consider sample sizes of 500 and 1000. The outcome of interest is the sum of the rewards from each stage, that is, $Y = R_1 + R_2$. Furthermore, we consider both a tree-type underlying optimal DTR and a non-tree-type one.

Treatment variables are set to take values in $\{0, 1, 2\}$ at each stage. For stage 1, we generate A_1 from the same model as A in Scenario 1, and generate stage 1

TABLE 2

Simulation results for Scenario 1 *with a single stage, three treatment options and twenty baseline covariates* (500 *replications*, $n = 500$). π *is the propensity score model*. (a) *and* (b) *indicate equal and varying penalties for treatment misclassification in the generative outcome model*. opt% *shows the empirical mean and standard deviation* (*SD*) *of the percentage of subjects correctly classified to their optimal treatments*. $\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$ *shows the empirical mean and SD of the expected counterfactual outcome obtained using the true outcome model and the estimated optimal regime*. $E\{Y^*(g^{\text{opt}})\} = 2$

π	Method	(a)		(b)	
		opt%	$\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$	opt%	$\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$
–	RG	66.7 (2.8)	1.34 (0.08)	63.5 (3.4)	1.30 (0.09)
	RF	51.6 (5.7)	1.03 (0.13)	62.7 (5.8)	1.37 (0.12)
Correct	OWL	36.3 (4.2)	0.73 (0.10)	38.4 (5.4)	0.63 (0.17)
	LZ	88.6 (9.4)	1.77 (0.20)	85.5 (0.11)	1.74 (0.21)
	ACWL-C_1	89.6 (5.0)	1.79 (0.11)	83.7 (6.0)	1.70 (0.13)
	ACWL-C_2	90.7 (4.6)	1.82 (0.11)	82.5 (6.2)	1.70 (0.13)
	T-RL	96.3 (4.1)	1.93 (0.10)	91.9 (6.7)	1.86 (0.13)
	T-RL-LH	96.8 (3.9)	1.94 (0.09)	92.8 (5.4)	1.89 (0.10)
Incorrect	OWL	32.6 (4.0)	0.65 (0.10)	34.5 (4.3)	0.56 (0.15)
	LZ	85.9 (12.6)	1.72 (0.26)	78.4 (15.4)	1.62 (0.30)
	ACWL-C_1	87.8 (5.5)	1.76 (0.12)	82.6 (6.3)	1.70 (0.13)
	ACWL-C_2	90.8 (4.3)	1.82 (0.10)	81.7 (6.3)	1.70 (0.13)
	T-RL	97.4 (2.4)	1.95 (0.07)	90.7 (7.7)	1.85 (0.14)
	T-RL-LH	97.9 (2.0)	1.96 (0.07)	92.0 (6.5)	1.87 (0.11)

RG, linear regression; RF, random forests; OWL, outcome weighted learning; LZ, method b Laber and Zhao (2015); ACWL-C_1 and ACWL-C_2, method by Tao and Wang (2017); T-RL, tree-based reinforcement learning; T-RL-LH, T-RL with one step lookahead.

reward as

$$R_1 = \exp\big[1.5 + 0.3X_4 - |1.5X_1 - 2|\{A_1 - g_1^{\text{opt}}(\mathbf{H}_1)\}^2\big] + \varepsilon_1,$$

with tree-type $g_1^{\text{opt}}(\mathbf{H}_1) = I(X_1 > -1)\{I(X_2 > -0.5) + I(X_2 > 0.5)\}$ or non-tree-type $g_1^{\text{opt}}(\mathbf{H}_1) = I(X_1 > -0.5)\{1 + I(X_1 + X_2 > 0)\}$, and $\varepsilon_1 \sim N(0, 1)$.

For stage 2, we have treatment $A_2 \sim \text{Multinomial}(\pi_{20}, \pi_{21}, \pi_{22})$, with $\pi_{20} = 1/\{1+\exp(0.2R_1-0.5)+\exp(0.5X_2)\}$, $\pi_{21} = \exp(0.2R_1-0.5)/\{1+\exp(0.2R_1-0.5)+\exp(0.5X_2)\}$ and $\pi_{22} = 1 - \pi_{20} - \pi_{21}$. We generate stage 2 reward as

$$R_2 = \exp\big[1.18 + 0.2X_2 - |1.5X_3 + 2|\{A_2 - g_2^{\text{opt}}(\mathbf{H}_2)\}^2\big] + \varepsilon_2,$$

with tree-type $g_2^{\text{opt}}(\mathbf{H}_2) = I(X_3 > -1)\{I(R_1 > 0) + I(R_1 > 2)\}$ or non-tree-type $g_2^{\text{opt}}(\mathbf{H}_2) = I(X_3 > -0.5)\{1 + I(X_3 + R_1 > 2)\}$, and $\varepsilon_2 \sim N(0, 1)$.

We apply the proposed T-RL algorithm with the modified pseudo-outcomes. For comparison, we apply Q-learning which uses the conditional mean models directly

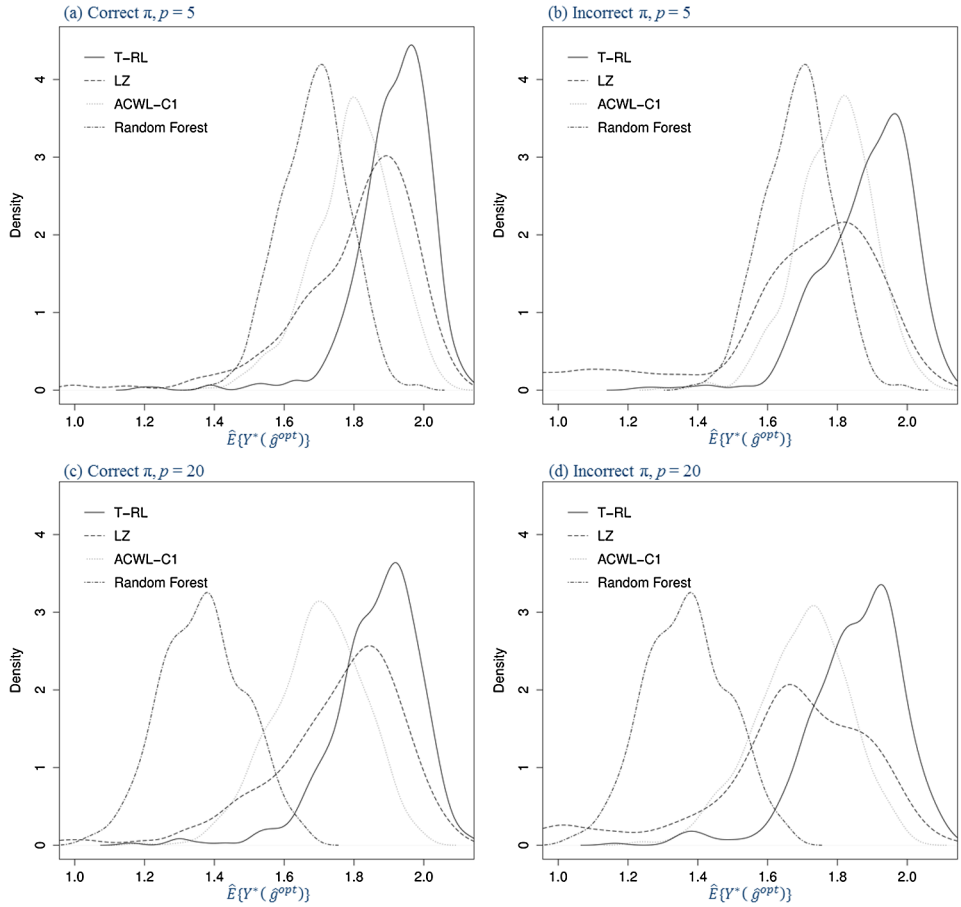

FIG. 2. *Density plots for the estimated counterfactual mean outcome in Scenario* 1 *with varying penalties for misclassification in the generative outcome model* (500 *replications*, $n = 500$). *The four panels are under correctly or incorrectly specified propensity model* (π) *and five or twenty baseline covariates* (p).

to infer the optimal regimes. We apply both the linear regression-based and random forests-based conditional mean models, denoted as Q-RG and Q-RF, respectively. We also apply the backward OWL (BOWL) method by Zhao et al. (2015) and the ACWL algorithm, both of which are implemented using the R package *rpart* for classification. In this scenario, we attempt to see how sample size and tree- or non-tree-type underlying DTRs affect the performances of various methods.

Results for Scenario 2 are shown in Table 3. ACWL and T-RL both work much better than Q-RG and BOWL in all settings. Q-RF is a competitive method only when the true optimal DTR is of a tree type, but it is consistently inferior to T-RL, likely due to its weakness in emphasizing prediction accuracy of the clinical response model instead of directly optimizing the decision rule. Given a tree-type

TABLE 3

Simulation results for Scenario 2 *with two stages and three treatment options at each stage* (500 *replications*). π *is the propensity score model*. opt% *shows the empirical mean and standard deviation* (*SD*) *of the percentage of subjects correctly classified to their optimal treatments*. $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ *shows the empirical mean and SD of the expected counterfactual outcome obtained using the true outcome model and the estimated optimal DTR*. $E\{Y^*(\mathbf{g}^{\text{opt}})\} = 8$

π	Method	Tree-type DTR				Non-tree-type DTR			
		$n = 500$		$n = 1000$		$n = 500$		$n = 1000$	
		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%**	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$
–	Q-RG	54.9 (3.0)	6.24 (0.16)	57.3 (2.5)	6.35 (0.13)	69.2 (4.4)	6.87 (0.17)	72.6 (3.5)	7.01 (0.13)
	Q-RF	84.0 (4.1)	7.53 (0.14)	92.1 (2.5)	7.80 (0.10)	74.4 (2.8)	7.11 (0.14)	77.5 (1.8)	7.25 (0.09)
Correct	BOWL	22.4 (5.1)	4.23 (0.38)	27.9 (5.7)	4.45 (0.43)	25.3 (6.0)	4.53 (0.42)	34.8 (7.0)	4.98 (0.47)
	ACWL-C_1	83.6 (5.9)	7.58 (0.18)	92.1 (4.4)	7.82 (0.11)	80.1 (6.1)	7.40 (0.18)	88.3 (3.4)	7.65 (0.11)
	ACWL-C_2	81.3 (6.6)	7.53 (0.20)	89.1 (5.7)	7.80 (0.12)	83.3 (5.5)	7.51 (0.16)	89.2 (3.0)	7.68 (0.11)
	T-RL	90.5 (7.0)	7.75 (0.22)	95.7 (3.6)	7.88 (0.11)	82.2 (4.6)	7.45 (0.14)	84.7 (2.9)	7.54 (0.11)
Incorrect	BOWL	16.4 (4.3)	4.20 (0.30)	16.5 (4.9)	4.29 (0.32)	16.3 (4.9)	4.29 (0.34)	17.9 (6.0)	4.56 (0.37)
	ACWL-C_1	80.9 (5.9)	7.55 (0.18)	89.1 (4.9)	7.80 (0.11)	73.4 (6.7)	7.30 (0.20)	81.0 (6.3)	7.56 (0.14)
	ACWL-C_2	80.4 (6.7)	7.54 (0.19)	86.4 (6.2)	7.76 (0.13)	79.8 (5.9)	7.46 (0.17)	86.3 (4.2)	7.66 (0.11)
	T-RL	90.2 (7.4)	7.74 (0.17)	93.9 (6.0)	7.87 (0.11)	82.2 (6.3)	7.47 (0.15)	84.8 (3.8)	7.55 (0.11)

Q-RG, Q-learning with linear regression; Q-RF, Q-learning with random forests; BOWL, backward outcome weighted learning; ACWL-C_1 and ACWL-C_2, method by Tao and Wang (2017); T-RL, tree-based reinforcement learning.

underlying DTR, T-RL has the best performance among all methods considered, regardless of the specification of the propensity score model. It has average opt% over 90% and $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ closest to the truth 8. The results are a bit more complex when the underlying DTR is non-tree-type. The tree-based methods of ACWL with CART and T-RL both have misspecified DTR structures and thus show less satisfactory performances. However, ACWL seems more robust to the DTR misspecification with ACWL-C_2 showing larger opt% and $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ in all settings except when sample size is 500 and π is misspecified, in which case T-RL's stronger robustness to propensity score misspecification dominates. With non-rectangular boundaries in a non-tree-type DTR, a split may not improve the counterfactual mean estimates at the current node but may achieve such a goal in the future nodes. T-RL, with a purity measure based on $E\{Y^*(g)\}$, will terminate the splitting as soon as the best split of the current node fails to improve the counterfactual mean outcome. In contrast, the misclassification error-based impurity measure in CART may continue the recursive partitioning as the best split may still reduce misclassification error without improving the counterfactual mean outcome at the current node. In other words, T-RL may be more myopic when it comes to non-tree-type DTRs.

Additional simulation results can be found in the Supplementary Material A [Tao, Wang and Almirall (2018a)], which leads to similar conclusions for these methods in comparison. In addition, the comparison of T-RL and the list-based method by Zhang et al. (2015) shows slightly better performance for T-RL given no cost information for measuring covariates. To implement the proposed method and the competing methods, the R codes and sample data can be found in the Supplementary Material B [Tao, Wang and Almirall (2018b)].

4. Application to substance abuse disorder data. We apply T-RL to the data of an observational study, where 2870 adolescents entered community-based substance abuse treatment programs, which are pooled from several adolescent treatment studies funded by the Center for Substance Abuse Treatment (CSAT) of the Substance Abuse and Mental Health Services Administration (SAMHSA). The measurements on individual characteristics and functioning are collected at baseline and at the end of three and six months. We use subscript values $t = 0, 1, 2$ to denote baseline, month three, and month six respectively.

Substance abuse treatments were given twice, first during months zero $\sim$ three, denoted as A_1 and second during months three $\sim$ six, denoted as A_2. At each stage, subjects were provided with one of the three options: no treatment, nonresidential treatment (outpatient only) and residential treatment (i.e., inpatient rehab) [Marlatt and Donovan (2005)], which we denote as 0, 1 and 2, respectively. At stage 1, 93% of the subjects received treatment, either residential (56%), or nonresidential (27%), while at stage 2, only 28% and 13% were treated residentially or non-residentially. We denote the baseline covariate vector for predicting the assignment of A_1 as $\mathbf{X}_1$ and the covariate history just before assigning A_2 as

$\mathbf{X}_1$ ($\mathbf{X}_1$ includes $\mathbf{X}_0$). The detailed list of variables used can be found in Almirall et al. (2012). The outcome of interest is the Substance Frequency Scale (SFS) collected during six $\sim$ nine months (mean and SD: 0.09 and 0.13), with higher values indicating increased frequency of substance use in terms of days used, days staying high most of the day, and days causing problems. We take $Y = -1 \times \text{SFS}$ so that higher values are more desired, making it consistent with our foregoing notation and method derivation. Missing data is imputed using IVEware [Raghunathan, Solenberger and Van Hoewyk (2002)].

We apply the T-RL algorithm to the data described above. Specifically, the covariate and treatment history just prior to stage 2 treatment is $\mathbf{H}_2 = (\mathbf{X}_1^\top, A_1)^\top$ and the number of treatment options at stage 2 is $K_2 = 3$. We fit a linear regression model for $\mu_{2,A_2}(\mathbf{H}_2)$ similar to (2.4) using Y as the outcome; all variables in $\mathbf{H}_2$ are included as interaction terms with A_2. For the propensity score $\pi_{2,A_2}(\mathbf{H}_2)$, we fit a multinomial logistic regression model including main effects of all variables in $\mathbf{H}_2$. We set the minimal node size to be 50 and maximum tree depth to be 5, and use a 10-fold CV to select λ, the minimum purity improvement for splitting. We repeat a similar procedure for stage 1 except that we have $\mathbf{H}_1 = \mathbf{X}_0$, $K_1 = 3$ and $\widehat{PO}_1' = Y + \hat{\mu}_{2,\hat{g}_2^{\text{opt}}}(\mathbf{H}_2) - \hat{\mu}_{2,A_2}(\mathbf{H}_2)$.

At stage 2, the variables in the estimated optimal regime are yearly substance dependence scale measured at the end of month three [sdsy3, median (range): 3 $(0-7)$], age [median (range): 16 $(12-25)$ years], and yearly substance problem scale measured at baseline [spsy0, median (range): 8 $(0-16)$]. At stage 1, the variables in the estimated optimal regime are emotional problem scale measured at baseline [eps7p0, median (range): 0.22 $(0-1)$], drug crime scale measured at baseline [dcs0, median (range): 0 $(0-5)$], and environmental risk scale measured at baseline [ers0, median (range): 35 $(0-77)$]. All these scale variables have higher values indicating more risk or problems. Specifically, the estimated optimal DTR is $\hat{\mathbf{g}}^{\text{opt}} = (\hat{g}_1^{\text{opt}}, \hat{g}_2^{\text{opt}})$, with

$$\hat{g}_1^{\text{opt}}(\mathbf{H}_1) = \begin{cases} \text{no treatment} & \text{if eps7p0} \le 0.286 \ \&\ \text{ers0} \le 46, \\ \text{non-residential} & \text{if eps7p0} \ge 0.286 \ \&\ \text{dcs0} \le 2, \\ \text{residential} & \text{otherwise,} \end{cases}$$

and

$$\hat{g}_2^{\text{opt}}(\mathbf{H}_2) = \begin{cases} \text{residential} \\ \qquad \text{if sdsy3} > 0, \ \text{or sdsy3} = 0 \ \&\ \text{age} \le 16 \ \&\ \text{spsy0} > 5, \\ \text{non-residential} \\ \qquad \text{otherwise.} \end{cases}$$

According to the estimated optimal DTR, at stage 1, subjects with fewer emotional problems and lower environmental risk do not need to be treated, while

those with more emotional problems but lower drug crime scale should be offered outpatient treatment only. At stage 2, all subjects should be treated. Those with higher yearly substance dependence as well as those with no yearly substance dependence but younger age and more yearly substance problems should receive residential treatment, that is, receiving treatment in rehab facilities. In contrast, subjects with older age or fewer yearly substance problems should be provided with outpatient treatment. The majority of subjects at both stages would benefit most from residential treatment. In our data, about 70% of the subjects at stage 1 have the estimated optimal treatment to be residential treatment and the number goes up to 85% at stage 2. Residential treatment is generally more intensive and subjects are in a safe and structured environment, which may explain why subjects with more substance, emotional or environmental problems would benefit more from this type of treatment. Existing studies have found a moderate level of evidence for the effectiveness of residential treatment for substance use disorders [Reif et al. (2014)]. Generally, outpatient programs allow subjects to return to their own environments during treatment. Subjects are encouraged to develop a strong support network of non-using peers and sponsors, and are expected to apply the lessons learned from outpatient treatment programs to their daily experiences [Gifford (2015)]. Nonetheless, subjects may respond sub-optimally to outpatient treatment (relative to residential treatment) if they have a larger network of peers that are using or at risk of using substances. Therefore, it may not be surprising that subjects with a lower environmental risk scale would benefit more from outpatient treatment.

5. Discussion. We have developed T-RL to identify optimal DTRs in a multi-stage multi-treatment setting, through a sequence of unsupervised decision trees with backward induction. T-RL enjoys the advantages of typical tree-based methods as being straightforward to understand and interpret, and capable of handling various types of data without distributional assumptions. T-RL can also handle multinomial or ordinal treatments by incorporating multiple treatment comparisons directly in the purity measure for node splitting, and thus works better than ACWL when the underlying optimal DTR is tree-type. Moreover, T-RL maintains the robust and efficient property of ACWL by virtue of the combination of robust semiparametric regression estimators with flexible machine learning methods, which is superior to IPW-based methods such as LZ. However, when the true optimal DTR is non-tree-type, ACWL has slightly more robust performances.

Several improvements and extensions can be explored in future studies. As shown by the simulation, the fixed-depth lookahead is costly and only brings moderate improvement. Alternatively, one can use embedded models to select splitting variables which also enjoys the lookahead feature [Zhu, Zeng and Kosorok (2015)], or consider other variants of lookahead methods [Elomaa and Malinen (2003), Esmeir and Markovitch (2004)]. The method by Zhu, Zeng and Kosorok (2015) enables progressively muting noise variables as one goes further down a

tree, which facilitates the modeling in high-dimensional sparse settings, and it also incorporates linear combination splitting rules, which may improve the identification of non-tree-type optimal DTRs. Furthermore, it is of great importance to explore how to handle continuous treatment options in the proposed T-RL framework. One way is to follow LZ to use a kernel smoother in the purity measure, which may suffer from the difficulty in selecting the optimal bandwidth. A simpler approach is to discretize the continuous treatments by certain quantiles and consider it as ordinal treatments, which may improve estimation stability and is also of practical interest as medical practitioners tend to prescribe treatments by several fixed levels instead of a continuous fashion.

Acknowledgments. The authors thank the grantees and their participants for agreeing to share their data to support the development of the statistical methodology.

SUPPLEMENTARY MATERIAL

Supplementary material A for article "Tree-based reinforcement learning for estimating optimal dynamic treatment regimes" (DOI: 10.1214/18-AOAS1137SUPPA; .pdf). Additional simulation results for the proposed method and competing methods.

Supplementary material B for article "Tree-based reinforcement learning for estimating optimal dynamic treatment regimes" (DOI: 10.1214/18-AOAS1137SUPPB; .zip). R codes and sample data to implement the proposed method.

REFERENCES

ALMIRALL, D., MCCAFFREY, D. F., GRIFFIN, B. A., RAMCHAND, R., YUEN, R. A. and MURPHY, S. A. (2012). Examining moderated effects of additional adolescent substance use treatment: Structural nested mean model estimation using inverse-weighted regression-with-residuals. Technical Report No. 12-121, Penn State Univ., Univiversity Park, PA.

BATHER, J. (2000). *Decision Theory*: *An Introduction to Dynamic Programming and Sequential Decisions*. Wiley, Chichester. MR1884596

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

CHAKRABORTY, B. and MOODIE, E. E. M. (2013). *Statistical Methods for Dynamic Treatment Regimes*: *Reinforcement Learning*, *Causal Inference*, *and Personalized Medicine*. Springer, New York. MR3112454

CHAKRABORTY, B. and MURPHY, S. (2014). Dynamic treatment regimes. *Annual Review of Statistics and Its Application* **1** 447–464.

CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Mach. Learn.* **20** 273–297.

ELOMAA, T. and MALINEN, T. (2003). On lookahead heuristics in decision tree learning. In *International Symposium on Methodologies for Intelligent Systems*. *Lecture Notes in Artificial Intelligence* **2871** 445–453. Springer, Heidelberg.

ESMEIR, S. and MARKOVITCH, S. (2004). Lookahead-based algorithms for anytime induction of decision trees. In *Proceedings of the Twenty-First International Conference on Machine Learning* 257–264. ACM, New York.

GIFFORD, S. (2015). Difference between outpatient and inpatient treatment programs. Psych Central. Retrieved on July 6, 2016, from http://psychcentral.com/lib/differences-between-outpatient-and-inpatient-treatment-programs.

HERNÁN, M. A., BRUMBACK, B. and ROBINS, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J. Amer. Statist. Assoc.* **96** 440–448. MR1939347

HSER, Y.-I., ANGLIN, M. D., GRELLA, C., LONGSHORE, D. and PRENDERGAST, M. L. (1997). Drug treatment careers A conceptual framework and existing research findings. *J. Subst. Abuse Treat.* **14** 543–558.

HUANG, X., CHOI, S., WANG, L. and THALL, P. F. (2015). Optimization of multi-stage dynamic treatment regimes utilizing accumulated data. *Stat. Med.* **34** 3423–3443. MR3412642

LABER, E. B. and ZHAO, Y. Q. (2015). Tree-based methods for individualized treatment regimes. *Biometrika* **102** 501–514. MR3394271

LAKKARAJU, H. and RUDIN, C. (2017). Learning cost-effective and interpretable treatment regimes. *Proceedings of Machine Learning Research* **54** 166–175.

MARLATT, G. A. and DONOVAN, D. M. (2005). *Relapse Prevention*: *Maintenance Strategies in the Treatment of Addictive Behaviors*. Guilford Press, New York, NY.

MCLELLAN, A. T., LEWIS, D. C., O'BRIEN, C. P. and KLEBER, H. D. (2000). Drug dependence, a chronic medical illness: Implications for treatment, insurance, and outcomes evaluation. *J. Am. Med. Dir. Assoc.* **284** 1689–1695.

MENARD, S. (2002). *Applied Logistic Regression Analysis*, 2nd ed. Sage, Thousand Oaks, CA.

MOODIE, E. E. M., CHAKRABORTY, B. and KRAMER, M. S. (2012). Q-learning for estimating optimal dynamic treatment rules from observational data. *Canad. J. Statist.* **40** 629–645. MR2998853

MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 331–366. MR1983752

MURPHY, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Stat. Med.* **24** 1455–1481. MR2137651

MURPHY, S. A., VAN DER LAAN, M. J. and ROBINS, J. M. (2001). Marginal mean models for dynamic regimes. *J. Amer. Statist. Assoc.* **96** 1410–1423. MR1946586

MURPHY, S. A., LYNCH, K. G., OSLIN, D., MCKAY, J. R. and TENHAVE, T. (2007). Developing adaptive treatment strategies in substance abuse research. *Drug Alcohol Depend.* **88** S24–S30.

MURTHY, S. and SALZBERG, S. (1995). Lookahead and pathology in decision tree induction. In *Proceedings of Fourteenth International Joint Conference on Artificial Intelligence* 1025–1031. Morgan Kaufmann, San Francisco, CA.

ORELLANA, L., ROTNITZKY, A. and ROBINS, J. M. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: Main content. *Int. J. Biostat.* **6** Art. 8, 49. MR2602551

RAGHUNATHAN, T. E., SOLENBERGER, P. and VAN HOEWYK, J. (2002). IVEware: Imputation and variance estimation software user guide. Survey Methodology Program, Univ. Michigan, Ann Arbor, MI.

REIF, S., GEORGE, P., BRAUDE, L., DOUGHERTY, R. H., DANIELS, A. S., GHOSE, S. S. and DELPHIN-RITTMON, M. E. (2014). Residential treatment for individuals with substance use disorders: Assessing the evidence. *Psychiatr. Serv.* (*Wash. D.C.*) **65** 301–312.

RIVEST, R. L. (1987). Learning decision lists. *Mach. Learn.* **2** 229–246.

ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512. MR0877758

ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412. MR1293185

ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, 69–117. Springer, New York. MR1601279

ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, 189–326. Springer, New York. MR2129402

ROBINS, J. M. and HERNÁN, M. A. (2009). Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis*, 553–599. CRC Press, Boca Raton, FL. MR1500133

ROTNITZKY, A., ROBINS, J. M. and SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Amer. Statist. Assoc.* **93** 1321–1339. MR1666631

SCHULTE, P. J., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2014). *Q*- and *A*-learning methods for estimating optimal dynamic treatment regimes. *Statist. Sci.* **29** 640–661. MR3300363

SUTTON, R. and BARTO, A. (1998). *Reinforcement Learning*: *An Introduction*. MIT Press, Cambridge.

TAO, Y. and WANG, L. (2017). Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. *Biometrics* **73** 145–155. MR3632360

TAO, Y., WANG, L. and ALMIRALL, D. (2018a). Supplement to "Tree-based reinforcement learning for estimating optimal dynamic treatment regimes." DOI:10.1214/18-AOAS1137SUPPA.

TAO, Y., WANG, L. and ALMIRALL, D. (2018b). Supplement to "Tree-based reinforcement learning for estimating optimal dynamic treatment regimes." DOI:10.1214/18-AOAS1137SUPPB.

THALL, P. F., WOOTEN, L. H., LOGOTHETIS, C. J., MILLIKAN, R. E. and TANNIR, N. M. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Stat. Med.* **26** 4687–4702. MR2413392

VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* **2** Art. 11, 40. MR2306500

WAGNER, E. H., AUSTIN, B. T., DAVIS, C., HINDMARSH, M., SCHAEFER, J. and BONOMI, A. (2001). Improving chronic illness care: Translating evidence into action. *Health Aff.* (*Millwood*, *Va.*) **20** 64–78.

WANG, L., ROTNITZKY, A., LIN, X., MILLIKAN, R. E. and THALL, P. F. (2012). Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *J. Amer. Statist. Assoc.* **107** 493–508. MR2980060

WATKINS, C. J. and DAYAN, P. (1992). Q-learning. *Mach. Learn.* **8** 279–292.

ZHANG, B., TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LABER, E. B. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat* **1** 103–114.

ZHANG, Y., LABER, E. B., TSIATIS, A. and DAVIDIAN, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* **71** 895–904. MR3436715

ZHANG, Y., LABER, E. B., TSIATIS, A. and DAVIDIAN, M. (2016). Interpretable dynamic treatment regimes. arXiv preprint arXiv:1606.01472.

ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** 1106–1118. MR3010898

ZHAO, Y.-Q., ZENG, D., LABER, E. B. and KOSOROK, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.* **110** 583–598. MR3367249

ZHOU, X., MAYER-HAMBLETT, N., KHAN, U. and KOSOROK, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *J. Amer. Statist. Assoc.* **112** 169–187. MR3646564

ZHU, R., ZENG, D. and KOSOROK, M. R. (2015). Reinforcement learning trees. *J. Amer. Statist. Assoc.* **110** 1770–1784. MR3449072

Y. TAO
L. WANG
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109
USA
E-MAIL: yebintao@umich.edu
luwang@umich.edu

D. ALMIRALL
INSTITUTE FOR SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48104
USA
E-MAIL: dalmiral@umich.edu