Published in final edited form as:
Stat Med. 2018 November 20; 37(26): 3776–3788. doi:10.1002/sim.7844.

Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens

Ying Liu[1], **Yuanjia Wang**[2], **Michael R. Kosorok**[3], **Yingqi Zhao**[4], and **Donglin Zeng**[3]

[1]Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI, USA

[2]Department of Biostatistics, Columbia University, New York City, NY, USA

[3]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[4]Public Health Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Abstract

Dynamic treatment regimens (DTRs) are sequential treatment decisions tailored by patient's evolving features and intermediate outcomes at each treatment stage. Patient heterogeneity and the complexity and chronicity of many diseases call for learning optimal DTRs that can best tailor treatment according to each individual's time-varying characteristics (eg, intermediate response over time). In this paper, we propose a robust and efficient approach referred to as Augmented Outcome-weighted Learning (AOL) to identify optimal DTRs from sequential multiple assignment randomized trials. We improve previously proposed outcome-weighted learning to allow for negative weights. Furthermore, to reduce the variability of weights for numeric stability and improve estimation accuracy, in AOL, we propose a robust augmentation to the weights by making use of predicted pseudooutcomes from regression models for Q-functions. We show that AOL still yields Fisher-consistent DTRs even if the regression models are misspecified and that an appropriate choice of the augmentation guarantees smaller stochastic errors in value function estimation for AOL than the previous outcome-weighted learning. Finally, we establish the convergence rates for AOL. The comparative advantage of AOL over existing methods is demonstrated through extensive simulation studies and an application to a sequential multiple assignment randomized trial for major depressive disorder.

Keywords

adaptive intervention; individualized treatment rule; machine learning; outcome-weighted learning; personalized medicine; Q-learning; SMARTs

1 | INTRODUCTION

Technology advances are revolutionizing medical research by collecting rich data from individual patient (eg, clinical assessments, genomic data, and electronic health records) for clinical researchers to meet the promise of individualized treatment and health care. The

Correspondence: Donglin Zeng, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. dzeng@email.unc.edu.

availability of comprehensive data sources provides new opportunities to deeply tailor treatment in the presence of patient heterogeneity and the complexity and chronicity of many diseases. Dynamic treatment regimens (DTRs),[1] also known as adaptive treatment strategies,[1] multistage treatment strategies,[2] or treatment policies,[3] are a sequence of treatment decisions adapted to the time-varying clinical status of a patient. Moreover, DTRs are necessary to treat complex chronic disorders such as major depressive disorder (MDD) when some patients fail to achieve remission with a first-line treatment.[4]

Sequential multiple assignment randomized trials (SMARTs), in which randomization is implemented at each treatment stage, have been advocated[5] to evaluate any DTR with causal interpretation. Using data collected from SMARTs, numerous methods have recently been developed to estimate optimal DTRs.[6–13] See also the works of Chakraborty and Moodie[14] and Kosorok and Moodie[15] for a detailed review of the current literature. Of all the methods, machine learning methods have received attention because of their robustness and computational advantages. For example, Q-learning[16] was used to analyze SMART data by Zhao et al[9] and Murphy et al.[17] In this learning algorithm, the optimal treatment at each stage is derived from a backward induction by maximizing the so-called Q-function (“Q” stands for “quality of action”), which is estimated via a regression model. To avoid model misspecification in Q-learning, Zhao et al[10] proposed outcome-weighted learning (OWL) to estimate the optimal treatment rules by directly optimizing the expected clinical outcome in a single-stage trial. They demonstrated in numerical studies that OWL outperforms Q-learning in small sample-size settings with many tailoring variables. Later, Zhao et al[18] generalized OWL to estimating optimal DTRs in a multiple-stage trial and demonstrated the superior performance to existing methods. However, in the aforementioned work,[18] since the weights at each stage of the estimation must be the optimal outcome increment in the future stages, only patients whose later treatments are optimal can be used for estimation. Consequently, a proportion of data have to be discarded from one stage to another in their backward learning algorithm, resulting in significant information loss and thus large variability of the estimated DTRs.

In this paper, we propose a hybrid approach, namely Augmented Outcome-weighted Learning (AOL), to integrate OWL and regression models for Q-functions for estimating the optimal DTRs. Similar to OWL, the proposed method relies on weighted machine learning algorithms in a backward induction. However, the weights used in AOL are constructed by augmenting optimal outcomes for all patients, including those whose later stage treatments are nonoptimal. The augmentation is obtained using prediction from the regression models for Q-functions. Thus, AOL performs augmented outcome-weighted learning using the regression models for Q-functions as augmentation.

There are several novel contributions in this work as compared with previous works.[10,18] First, for single-stage randomized trials, AOL generalizes OWL to allow for negative outcome values instead of adding an arbitrarily large constant, which may lead to numeric instability. Second, by using weights based on residuals after removing prognostic effects that are obtained from the observed outcomes, AOL reduces the variability of weights in OWL to achieve less variable DTR estimation. Third, AOL simultaneously takes advantage of the robustness of nonparametric OWL and makes use of model-based approaches to utilize data from all subjects. Fourth, AOL is theoretically shown to yield the same asymptotic bias as OWL but smaller stochastic variability because of a better weighting scheme and thus guarantees efficiency gain. Moreover, AOL is proved to yield the correct optimal DTRs even if the regression models assumed in the augmentation are incorrect and thus maintains the robustness of OWL.

The rest of this paper is organized as follows. In Section 2, we review some concepts for DTR, Q-learning, and OWL and introduce AOL for single-stage and multiple-stage studies. The last part of Section 2 presents theoretical properties of AOL. In particular, we provide stochastic error bounds for AOL and demonstrate its smaller stochastic variability when compared with OWL; we further derive a fast convergence rate for AOL. Section 3 shows the results of extensive simulation studies to examine the performance of AOL compared with Q-learning and OWL. In Section 4, we present real data analysis results from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial[4] for MDD. Lastly, we conclude with a few remarks in Section 5.

2 | METHODOLOGIES

2.1 | Dynamic treatment regimes and outcome weighted learning

We start by introducing notation for a K-stage DTR. For $k = 1, 2, \dots, K$, denote X_k as the observed subject-specific tailoring variables collected just prior to the treatment assignment at stage k. Denote A_k as the treatment assignment taking values in $\{-1, 1\}$, and R_k as the clinical outcome (also known as the “reward”) post the kth-stage treatment. Larger rewards may correspond to better functioning or fewer symptoms depending on the clinical setting. A DTR is a sequence of decision functions, $\mathscr{D} = (\mathscr{D}_1, \mathscr{D}_2, \dots, \mathscr{D}_K)$, where $\mathscr{D}_K$ maps the domain of patient health history information, $H_k = (X_1, A_1, R_1, \dots, A_{k-1}, R_{k-1}, X_k)$, to the treatment choices in $\{-1, 1\}$. Corresponding to each $\mathscr{D}$, a value function, denoted by $\mathscr{V}(\mathscr{D})$, is defined as the expected reward given that the treatment assignments follow regimen $\mathscr{D}$.[12] Mathematically, $\mathscr{V}(\mathscr{D}) = E_{\mathscr{D}}\left[\sum_{k=1}^{K} R_k\right] = \int \sum_{k=1}^{K} R_k dP_{\mathscr{D}}$, where $\mathscr{P}_{\mathscr{D}}$ is the probability measure generated by random variables $(X_1, A_1, R_1, \dots, X_K, A_K, R_K)$ given that $A_k = \mathscr{D}_k(H_k)$ and $E_{\mathscr{D}}$ is the expectation with respect to this measure. Hence, the goal of personalized DTRs is to find the optimal DTRs that maximize the value function.

To evaluate the value function of a DTR in a SMART, a potential outcome framework in causal inference literature is used. The potential outcome in our context is defined as the outcome of a subject had he or she followed a particular treatment regimen, possibly different from the observed regimen in the actual trial. Several assumptions are required to infer the value function of a DTR, including the standard stable unit treatment value assumption and the no unmeasured confounders assumption.[6,19] In a SMART, the no unmeasured confounders assumption is automatically satisfied because of the virtue of sequential randomization. Furthermore, we need the following positivity assumption: let $\pi_k(a, h)$ denote the treatment assignment probability, $P(A_k = a|H_k = h)$, which is given by design so known to investigators in a SMART. We assume that, for $k = 1, \dots, K$ and any $a \in \{-1, 1\}$ and h_k in the support of H_k, $\pi_k(a, h_k) = P(A_k = a|H_k = h_k) \in [c, \tilde{c}]$, where $0 < c \le \tilde{c} <$

1 are two constants. That is, the positivity assumption requires that each DTR has a positive chance of being observed.

Under these assumptions, if we let P denote the probability measure generated by (X_k,A_k, R_k) for k = 1, … , K, then according to the work of Qian and Murphy,[12] it can be shown that $P_{\mathscr{D}}$ is dominated by P and

$$\mathscr{V}(\mathscr{D}) = E\left[\frac{\prod_{k=1}^{K} I\left(A_k = \mathscr{D}_k\left(H_k\right)\right)\left(\sum_{k=1}^{K} R_k\right)}{\prod_{k=1}^{K} \pi_k\left(A_k, H_k\right)}\right]. \quad (1)$$

Consequently, the goal is to find the optimal treatment rule $\mathscr{D}^* = \left(\mathscr{D}_1^*, \ldots, \mathscr{D}_K^*\right)$ that maximizes the above expectation. Note that $\mathscr{D}_k$ is usually given as the sign of some decision function f_k. Without confusion, we sometimes express the value function as $\mathscr{V}\left(f_1, \ldots, f_K\right)$ to emphasize its dependence on the decision functions.

Denote data collected from n i.i.d. subjects in a SMART at stage k as (A_{ik}, H_{ik}, R_{ik}) for i = 1, … , n, k = 1, … , K. Recently, outcome-weighted learning (Zhao et al), abbreviated as OWL, was proposed to estimate the optimal treatment regimes. Specifically, Zhao et al proposed a backward induction to implement OWL, where at stage k, they used only the subjects who followed the estimated optimal treatment regimens after stage k in the optimization algorithm. That is, the optimal rule $\mathscr{D}_k\left(H_k\right) = \operatorname{sign}\left(f_k\left(H_k\right)\right)$ solves a weighted support vector machine problem

$$\min_{f \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} \phi\left(A_{ik} f_k\left(H_{ik}\right)\right) \frac{\sum_{j=k}^{K} R_{ij}}{\pi_{ik}} \frac{I\left(A_{i,k+1} = \widehat{\mathscr{D}}_{k+1}\left(H_{i,k+1}\right), \ldots, A_{iK} = \widehat{\mathscr{D}}_K\left(H_{iK}\right)\right)}{\prod_{j>k} \pi_{ij}} + \lambda_n \| f_k \|^2 , \quad (2)$$

where $\varnothing(x)$ = max(0, 1 − x) is the hinge loss $\pi_{ij} = \pi_j\left(A_{ij}, H_{ij}\right)$, $\widehat{\mathscr{D}}_j\left(H_{ij}\right)$ is the estimated optimal rule at stage j from the backward learning algorithm, and $\|f\|$ is some norm defined in a given metric space, $\mathscr{H}$, usually a reproducing kernel Hilbert space (RKHS), for f. However, as discussed before, the use of only those subjects who followed the optimal regimen in future stages may result in information loss, especially when K is not small. Furthermore, the work of Zhao et al[10] suggests to subtract a constant from R_{ik} to ensure a positive weight in the optimization algorithm, where the choice of constant is arbitrary and can be numerically influential in the above optimization.

2.2 | AOL with $K = 1$ stage

We first describe the proposed method, namely AOL, under the single-stage randomized trial setting ($K = 1$). The main idea of AOL is to improve OWL by replacing R_1 in (1) by some surrogate variable, which should give the same optimal decision rule but with less variability in the empirical estimation of (2).

Note that, for any integrable function $s(H_1)$ and for $\mathscr{D}(H_1) = \text{sign}(f_k(H_k))$, it holds that

$$\begin{aligned}\mathscr{V}(\mathscr{D}) &= E\left[\frac{I(A_1 f(H_1) > 0)R_1}{\pi_1(A_1, H_1)}\right]\\ &= E\left[\frac{I(A_1 f(H_1) > 0)(R_1 - s(H_1))}{\pi_1(A_1, H_1)}\right] + E[s(H_1)]\\ &= E\left[\frac{|R_1 - s(H_1)|}{\pi_1(A_1, H_1)} I(A_1 \text{sign}(R_1 - s(H_1)) f(H_1) > 0)\right] + E[s(H_1)] - E\left[\frac{(R_1 - s(H_1))^-}{\pi_1(A_1, H_1)}\right],\end{aligned}$$

where $x^- = -\min(0, x)$. Therefore, maximizing $V(D)$ is equivalent to maximizing

$$E\left[\frac{|R_1 - s(H_1)|}{\pi_1(A_1, H_1)} I(A_1 \text{sign}(R_1 - s(H_1)) f(H_1) > 0)\right].$$

This suggests that, if we choose a surrogate variable, $\tilde{R}_1 = R_1 - s(H_1)$, to replace R_1 and solve a similar problem to (2), where the weights are changed to $|\tilde{R}_{i1}|/\pi_{i1}$ and the class labels become $A_{i1}\text{sign}(\tilde{R}_{i1})$, then we expect to still obtain a consistent estimator of the optimal DTR.

Specifically, the proposed AOL for the $K = 1$ stage consists of the following two steps.

<u>Step 1</u>. Use data (R_{i1}, H_{i1}) to obtain an estimator $\hat{s}(H_1) = \hat{\gamma}_0 + \hat{\gamma}_1^T H_1$ by fitting a least squares regression or a penalized least squares regression if H_{i1} is high dimensional.

<u>Step 2</u>. Obtain $\tilde{R}_{i1} = R_{i1} - \hat{s}(H_{i1})$ for each subject and fit a weighted support vector machine (SVM) to estimate the decision function f_1, where the weights are $|\tilde{R}_{i1}|/\pi_{i1}$ and the class labels are $A_{i1}\text{sign}(\tilde{R}_{i1})$. That is, the estimated decision function, denoted by $\hat{f}_1$, minimizes

$$n^{-1}\sum_{i=1}^{n}\frac{|\tilde{R}_{i1}|}{\pi_{i1}}\phi(A_{i1}\text{sign}(\tilde{R}_{i1})f_1(H_{i1})) + \lambda_n \parallel f_1 \parallel .$$

The function class for f_1 is from an RKHS with either a linear kernel or a Gaussian kernel, which are the most popular choices in practice, although the proposed method can be applied with any kernels. Computationally, this minimization can be carried out using quadratic programming.[20] Finally, the optimal DTR, $\mathscr{D}_1^*$, is estimated as $\widehat{\mathscr{D}}_1(H_1) = \mathrm{sign}\left(\hat{f}_1(H_1)\right)$.

Remark 1. A heuristic interpretation of AOL is the following: first, learning DTR is essential to learn the qualitative interaction between A_1 and H_1, so the removal of any main effects $s(H_1)$ from R_1 has no influence; second, for a subject with large observed value of $|R_1 - s(H_1)|$, the above maximization implies that the optimal treatment assignment should be likely to remain the same as the actual treatment he/she is observed to receive in a trial if $R_1 - s(H_1)$ is positive but should be the opposite if negative. Furthermore, there are intuitive advantages to use $\tilde{R}_1$ to replace R_1 and use $\left|\tilde{R}_1\right|/\pi_1(A_1, H_1)$ as the new weight. When $s(H_1)$ is chosen appropriately, the resulting $\tilde{R}_1$ is less variable, so we expect that it may lead to a less variable DTR estimator using empirical observations. Moreover, since the proposed new weights are nonnegative, this guarantees a convex optimization problem when solving (2). In contrast, in the original OWL, when the weights in (2) are negative, they suggested subtracting an arbitrarily small constant from the weights to make it positive. This shifting of negative weights has been demonstrated to be unstable in numerical studies.

2.3 | AOL with $K = 2$ stages

Next, we consider $K = 2$. Because DTRs aim to maximize the expected cumulative rewards across all stages, the optimal treatment decision rule at the current stage must depend on subsequent decision rules and future clinical outcomes or rewards under those rules. This observation motivates us to use a backward procedure similar to the backward induction in Q-learning and OWL in the work of Zhao et al.[18] To estimate the optimal treatment rule at stage 2, AOL has the same two steps as in Section 2.2.

Step 2–1. Use data (R_{i2}, H_{i2}) to obtain an estimator $\hat{s}_2(H_2) = \hat{\gamma}_0 + \hat{\gamma}_1^T H_2$ by fitting a least squares regression or a penalized least squares regression if H_{i2} is high dimensional.

Step 2–2. Obtain $\tilde{R}_{i2} = R_{i2} - \hat{s}_2(H_{i2})$ for each subject and fit a weighted SVM to estimate the decision function f_2, where the weights are $\left|\tilde{R}_{i2}\right|/\pi_{i2}$, and the class labels are $A_{i2}\mathrm{sign}\left(\tilde{R}_{i2}\right)$. That is, the estimated function, denoted by $\hat{f}_2$, minimizes

$$n^{-1}\sum_{i=1}^{n}\frac{\left|\tilde{R}_{i2}\right|}{\pi_{i2}}\phi(A_{i2}\mathrm{sign}(\tilde{R}_{i2})f(H_{i2})) + \lambda_n \parallel f \parallel .$$

Thus, the estimated optimal DTR at stage 2 is given by $\widehat{\mathscr{D}}_2(H_2) = \mathrm{sign}\left(\hat{f}_2(H_2)\right)$.

Now, we consider the estimation of the optimal stage 1 treatment rule. For this purpose, a key outcome variable is the so-called Q-function, denoted by Q_2, which is the future reward increment at future stages if a subject is assigned to the optimal treatment in those stages. If Q_2 were observed for each subject, then the optimal treatment rule at stage 1 would be estimated using OWL with $R_1 + Q_2$ as the outcome part of the weight. For the subjects whose treatment assignments at stage 2 are the same as the optimal treatment rule $\mathscr{D}_2^*$, it is clear that $Q_2 = R_2$, and thus, their weights are observed; however, for subjects whose treatment assignments at stage 2 are not optimal (ie, not the same as $\mathscr{D}_2^*$), Q_2 is not observed. Moreover, OWL uses only those subjects whose Q_2's are observed and multiplies by the inverse probability of treatment assignment.

However, if we treat missing Q_2 as a missing data problem, it is well known that the use of only complete data for estimation may not be the most efficient method; instead, one can use auxiliary information prior to stage 2, namely H_2, to predict Q_2 through augmentation for those subjects with missing Q_2 (ie, for those subjects whose treatment assignments at stage 2 are not the same as the optimal treatment rule $\mathscr{D}_2^*$). Define $m_{22}(H_2)$ as an approximation to the optimal reward increment for subjects who receive nonoptimal treatment at stage 2. Following the missing data literature,[21] such an augmented Q_2 can be defined as

$$\frac{I\left(A_2=\mathscr{D}_2^*(H_2)\right)}{\pi_2\left(A_2,H_2\right)}R_2-\left\{\frac{I\left(A_2=\mathscr{D}_2^*(H_2)\right)}{\pi_2\left(A_2,H_2\right)}-1\right\}m_{22}\left(H_2\right).$$

Ideally, we want to choose $m_{22}(H_2)$ as close as possible to $E[Q_2|H_2]$; however, in practice, because the latter is unknown, and H_2 can be high dimensional, we will estimate $m_{22}(H_2)$ as a linear function of H_2 using a weighed least squares regression for subjects who are treated optimally in stage 2 as described below. To estimate the optimal stage 1 treatment rule, AOL has the following steps.

Step 1–1. Recall $\widehat{\mathscr{D}}_2\left(H_2\right) = \mathrm{sign}\left(\hat{f}_2\left(H_2\right)\right)$. Estimate $m_{22}(H_2) = \beta_0 + \beta^T H_2$ by a weighted least squares regression minimizing

$$n^{-1}\sum_{i=1}^{n}\frac{I\left(A_{i2}=\widehat{\mathscr{D}}_2\left(H_{i2}\right)\right)}{\pi_{i2}}\frac{1-\pi_{i2}}{\pi_{i2}}\left(R_{i2}-m_{22}\left(H_{i2}\right)\right)^2,$$

and denote the resulting estimator as $\hat{m}_{22}$.

Step 1–2. For subject i, compute

$$\hat{Q}_{i2} = \frac{I\left(A_{i2}=\hat{D}_2\left(H_{i2}\right)\right)}{\pi_{i2}}R_{i2}-\left\{\frac{I\left(A_{i2}=\hat{D}_2\left(H_{i2}\right)\right)}{\pi_{i2}}-1\right\}\hat{m}_{22}\left(H_{i2}\right).$$

Step 1–3. Obtain an estimator $\hat{s}_1\left(H_1\right)$ for $s_1(H_1) = \alpha_0+\alpha^T H_1$ using a least squares regression that minimizes

$$\sum_{i=1}^{n}\left(R_{i1}+\hat{Q}_{i2}-s_1\left(H_{i1}\right)\right)^2$$

and denote $\tilde{R}_{i1} = R_{i1}+\hat{Q}_{i2}-\hat{s}_1\left(H_{i1}\right)$.

Step 1–4. Finally, obtain $\hat{f}_1$ by fitting a weighted SVM with weights $\left|\tilde{R}_{i1}\right|/\pi_{i1}$ and class labels $A_{i1}\mathrm{sign}\left(\tilde{R}_{i1}\right)$. The optimal DTR at stage 1 is then $\hat{\mathscr{D}}_1\left(H_1\right) = \mathrm{sign}\left(\hat{f}_1\left(H_1\right)\right)$.

Note that the last two steps (Steps 1–3 and 1–4) essentially repeat the same procedure as in the $K = 1$ stage except that the outcome is the augmented outcome variable $R_{i1}+\hat{Q}_{i2}$ As a remark, when H_{i2} or H_{i1} is of high dimension, we recommend that a penalized least squares regression such as Lasso be used in Step 1–1 or Step 1–3 in practice.

Remark 2. The key idea of our proposed approach for two-stage problem is to use prediction models for the Q-function at stage 2 to "impute" the future reward increments for the subjects whose actual treatments received in the second stage are not the optimal because their the observed outcomes cannot be used to estimate the optimal future reward increments. The missingness mechanism is due to the randomization of the treatments in stage 2; thus, it is completely known. The proposed augmented weights in stage 1 are guaranteed to yield the correct optimal treatment rules. Furthermore, if the "imputation" is sufficiently close to the underlying true model, we expect to obtain better accuracy in finding the optimal rule because of using more observations.

2.4 | Generalization to more than 2 stages

When there are more than two stages, the same backward learning as in $K = 2$ can be applied, but the augmentation for those subjects with missing future optimal reward increments becomes more complex. First, to estimate the optimal treatment rule at stage K, we perform the same stage 2 steps as AOL with $K = 2$ (ie, Steps 2–1 and 2–2 in Section 2.3) but with (R_2, A_2, H_2) replaced by (R_K, A_K, H_K). Denote the resulting estimated decision function at this stage as $\hat{f}_K\left(H_K\right)$ and denote the corresponding treatment rule as $\hat{\mathscr{D}}_K\left(H_K\right) = \mathrm{sign}\left(\hat{f}_K\left(H_K\right)\right)$.

We then continue to estimate the optimal treatment rules at stage $K - 1$, $K - 2$, … in turn. Specifically, to estimate the optimal $(k - 1)$th-stage treatment rule, we let $M_{i,k}^{k-1} = 1$, and for $j \geq k$, let $M_{ik}^{j} = I\left(A_{ik} = \mathscr{D}_k^*\left(H_{ik}\right), \ldots, A_{ij} = \mathscr{D}_j^*\left(H_{ij}\right)\right)$ denote whether subject i follows the optimal treatment regimens from stage k to j. From the theory of Robins[8], also seen in Tsiatis[21] and Zhang et al[22], Q_{ik}, the optimal reward increment for patient i if she/he follows the estimated optimal rule from stage k to K, has the following expression:

$$Q_{ik} = \frac{M_{ik}^{K}\left(R_{ik}+\ldots+R_{iK}\right)}{\prod_{l=k}^{K}\pi_{il}} - \sum_{j=k}^{K}\left\{\frac{M_{ik}^{j-1}}{\prod_{l=k}^{j-1}\pi_{il}}\left\{\frac{I\left(A_{ij}=\mathscr{D}_j^*\left(H_{ij}\right)\right)}{\pi_{ij}}-1\right\}m_{kj}\left(H_{ij}\right)\right\},$$

where $m_{kj}(H_{ij})$ is the optimal reward increment for subjects who receive optimal treatments up to stage $(j - 1)$, ie, $E\left(Q_{ik} \middle| H_{ij}, M_{ik}^{j-1} = 1\right)$.

To implement AOL, at stage $k - 1$, assume that we have already obtained the estimated optimal rules after this stage, denoted by $\widehat{\mathscr{D}}_k, \ldots \widehat{\mathscr{D}}_K$. Define

$$\widehat{M}_{ik}^{j} = I\left(A_{ik} = \widehat{\mathscr{D}}_k(H_{ik}), \ldots, A_{ij} = \widehat{\mathscr{D}}_j(H_{ij})\right).$$

Then, the augmentation term for Q_{ik} is estimated by

$$\frac{\widehat{M}_{ik}^{K}(R_{ik} + \ldots + R_{iK})}{\prod_{l=k}^{K} \pi_{il}} - \sum_{j=k}^{K} \left\{ \frac{\widehat{M}_{ik}^{j-1}}{\prod_{l=k}^{j-1} \pi_{il}} \left\{ \frac{I\left\{A_{ij} = \widehat{\mathscr{D}}_j\left(H_{ij}\right)\right\}}{\pi_{ij}} - 1 \right\} \widehat{m}_{kj}\left(H_{ij}\right) \right\}, \quad (3)$$

where $\widehat{m}_{kj}\left(H_j\right)$ is estimated as a linear function of H_j by the weighted least squares

$$n^{-1} \sum_{i=1}^{n} \frac{\widehat{M}_{ik}^{K}}{\prod_{l=k}^{K} \pi_{ik}} \frac{1 - \pi_{ij}}{\prod_{k \le l \le j} \pi_{il}} \left(\sum_{l=k}^{K} R_{il} - m_{kj}\left(H_{ij}\right) \right)^2 .$$

We define $\widetilde{R}_{i,k-1} = R_{i,k-1} + \widehat{Q}_{i,k} - \widehat{s}_{k-1}\left(H_{k-1}\right)$, where $\widehat{s}_{k-1}\left(H_{k-1}\right)$ is estimated via a least squares regression that minimizes $\sum_{i=1}^{n} \left\{R_{i,k-1} + \widehat{Q}_{ik} - s_{k-1}\left(H_{i,k-1}\right)\right\}^2$ for $s_{k-1}\left(H_{k-1}\right) = \alpha_0 + \alpha^T H_{k-1}$. Then, we will estimate $\widehat{f}_{k-1}$. by fitng a weighted SVM with weights $\left|\widetilde{R}_{i,k-1}\right| / \pi_{i,k-1}$ and class labels $A_{i,k-1}\mathrm{sign}\left(\widetilde{R}_{i,k-1}\right)$, ie, $\widehat{f}_{k-1}$ minimizes

$$n^{-1} \sum_{i=1}^{n} \frac{\left|\widetilde{R}_{i,k-1}\right|}{\pi_{i,k-1}} \phi(A_{i,k-1}\mathrm{sign}(\widetilde{R}_{i,k-1}) f_{k-1}(H_{i,k-1})) + \lambda_n \parallel f_{k-1} \parallel .$$

One important fact for AOL is that the estimated treatment rules are invariant even if we shift R_k by any constant c_k for $k = 1, \ldots, K$. This is because under constant translation, m_{kj} will be shifted by $\sum_{l=k}^{K} c_l$ so $\widehat{Q}_{ik}$ becomes $\widehat{Q}_{ik} + \sum_{l=k}^{K} c_l$. Therefore, $\widetilde{R}_{i,k-1}$ which is the residual after regressing $R_{i,k-1} + c_{k-1} + \widehat{Q}_{ik}$ on 1 and $H_{i,k}-1$, remains unchanged, so the estimated treatment rule is the same as before. Finally, when H_j's dimension is large, a penalized least square regression such as Lasso is recommended in the above procedure to obtain $\widehat{m}_{kj}\left(H_j\right)$.

2.5 | Software

We provide an R-package “DTRlearn” https://cran.r-project.org/web/packages/DTRlearn/index.html on CRAN for the single- and multiple-stage implementation of our proposed method (AOL) and Q-learning and O-learning as compared in the following simulation results and real data implementation.

2.6 | Summary of theoretical results

In the supplementary material, we provide theoretical justification for the proposed methods. Theorem A.1 provides an error bound for single stage AOL. We formally prove that using this new surrogate weight on the basis of the residuals of R_1, the value loss due to using the estimated treatment rule $\hat{f}_1$ has the same deterministic error bound as using the original R_1; however, the error bound due to data randomness is smaller. In this sense, the value function for AOL has the same approximation bias as OWL but a smaller stochastic error asymptotically. Thus, AOL requires fewer observations than OWL to achieve a similar performance.

Theorem A.2 in the supplementary material provides the improved risk bound for multiple stage AOL. We formally show that the above data augmentation method using a surrogate function $m_{22}(H_2)$ for subjects with missing Q_2 values will not increase the approximation bias of the value function estimation based on $\hat{f}_1$; furthermore, we show that compared with OWL, the estimation of $m_{22}(H_2)$ from a weighted least squares in Step 1–1 always leads to a smaller stochastic error bound of the value function estimation. Finally, Theorem A.3 gives a fast convergence rate of AOL under some regularity conditions.

The key idea behind the proofs is to decompose the value function associated with the estimated DTR into two parts: one is the bias due to considering the decision functions f_k at each stage from an RKHS; the other part is the stochastic error due to both the empirical approximation of the value function in terms of the augmented weights in the optimization. The former can be characterized in terms of the richness of the Hilbert space, whereas the latter depends on both the complexity of the function classes in the Hilbert space and, more importantly, the variability of the weights used in our proposed weighted SVM methods. The less variable the weights are, the smaller the stochastic error is. Therefore, the proposed method, which relies on the augmentation, tends to bring more information to reduce the variability in the weights.

3 | SIMULATION STUDIES

We conducted extensive simulation studies to compare AOL with existing approaches using the value function (reward) of the estimated optimal treatment rules. We compared three methods: (i) Q-learning based on linear regression models with a Lasso penalty; (ii) OWL as in the works of Zhao et al[10,18]; (iii) AOL as described in Section 2.

3.1 | Simulation settings

We simulated single-stage, two-stage, and four-stage randomized trials. In this section, we present the results of four-stage settings. In the supplementary material, we provide additional results of the single-stage (Section B.1) and two-stage (Section B.2) settings.

In the first four-stage scenario, we simulated a vector of baseline feature variables of dimension 20, $X_1 = (X_{1,1}, \dots, X_{1,20})$, from a multivariate normal distribution, where the first 10 variables had a pairwise correlation of 0.2, the remaining 10 variables were uncorrelated among one another and were also independent of the first 10 X's, and the variance for $X_{1,j}$ was 1 for $j = 1, \dots, 20$. The reward functions were generated as follows:

$$R_1 = X_{1,1}A_1 + \mathcal{N}(0,1);\ R_2 = \left(R_1 + X_{1,2}^2 + X_{1,3}^2 - 0.8\right)A_2 + \mathcal{N}(0,1);$$
$$R_3 = 2\left(R_2 + X_{1,4}\right)A_3 + X_{1,5}^2 + X_{1,6} + \mathcal{N}(0,1);\ R_4 = \left(R_3 - 0.5\right)A_4 + \mathcal{N}(0,1).$$

The randomization probabilities of treatment assignment at each stage were allowed to depend on the feature variables through

$$P\left(A_1 = 1 \middle| H_1\right) = \frac{1}{1 + \exp\left(-0.5X_{1,1}\right)};P\left(A_2 = 1 \middle| H_2\right) = \frac{1}{1 + \exp\left(0.1R_1\right)};P\left(A_3 = 1 \middle| H_3\right) = \frac{1}{1 + \exp\left(0.2X_{1,3}\right)};$$
$$P\left(A_4 = 1 \middle| H_4\right) = \frac{1}{1 + \exp\left(0.2X_{1,4}\right)}.$$

Patient's health history information matrix at stage k, H_k, was defined recursively by $(H_{k-1}, A_{k-1}, A_{k-1}H_{k-1}, R_{k-1})$, and at the first stage, it only contains the baseline feature variables, ie, $H_1 = X_1$. Therefore, there were $p = 20$ features for OWL and AOL in the first stage, $2p + 2$ for the second stage, and $8p + 14$ variables for the fourth stage. To handle high dimensionality of the feature space, especially when k increases, weighted least squares with a Lasso penalty was used to estimate $\hat{m}_{kj}$, and ordinary least squares with Lasso penalty was used to estimate $\hat{s}_k$. When estimating conditional expectations in Q-learning, (H_k, A_k) was included in the linear regression models (the number of predictors approximately doubles compared to OWL and AOL), and a Lasso penalty was imposed for better fitting.

In the second four-stage scenario, we imitated a real-world scenario of treating chronic mental disorders,[4] where the patient population consisted of several subgroups that respond to DTRs differently. However, because of unknown and complex treatment mechanisms, instead of directly observing subgroup memberships, only group-informative feature variables (such as clinical symptomatology measures or neuroimaging biomarkers) were observed. Specifically, we created 10 subgroups of equal size and let $G = 1, \dots, 10$ denote group. For group $G = l$, the optimal DTRs across 4 stages were

$$A_{jl}^{*} = 2([l/(2j-1)]\mathrm{mod}2) - 1,\ j = 1,2,3,4.$$

To simulate data from a SMART, we randomly generated their treatment assignments with equal probabilities at each stage, and for a subject in group $G = l$, we generated their reward outcomes as $R_1 = R_2 = R_3 = 0$ and $R_4 = \sum_{j=1}^{4} A_j A_{jl}^{*} + N(0,1)$. Furthermore, we generated potentially group-informative baseline feature variables, $X_1 = (X_{1,1}, \dots, X_{1,30})$, from a multivariate normal distribution with means depending on group membership: for patients in group $G = l$, the center of $X_{1,1}, \dots, X_{1,10}$ had a group-specific mean value μ_l, which was generated from $\mu_l \sim N(0, 5)$, while the means of the remaining feature variables, $X_{1,11}, \dots,$ $X_{1,30}$, were all zero. The first 10 features had a pairwise correlation of 0.2, and the remaining 20 variables were uncorrelated. Therefore, only $X_{1,1}, \dots, X_{1,10}$ were informative of the patient subgroup (and thus the optimal DTRs), and the remaining variables were noise. Since the group membership was not observed, the available data for our analysis consisted of $(X_1, A_1, A_2, A_3, A_4, R_4)$. For each data set, we applied Q-learning, OWL, and AOL to estimate the optimal rule. For OWL, we implemented the same algorithm as in the work of Zhao et al,[18] and the minimal value of the reward outcome was subtracted from the outcome to ensure the weights to be positive. In this setting, the clusters and optimal decision boundaries are fixed for each replication but different across replications. Thus, the results do not depend on the specific cluster arrangements. The decision boundaries are not explicitly determined by the observed predictors, but they are determined by the underlying latent classes, which confer information from the observed predictors.

At each stage k, H_k contained baseline feature variables X_1, previous stage treatment assignments, and products between X_1 and previous stage treatments. We varied sample sizes in the simulations. Cross-validation was used to choose the tuning parameter in Lasso regressions and was used to choose the tuning parameter of the SVM (from a grid of 2^{-15} to 2^{15}). The linear kernel was used for OWL and AOL. To compare all the methods, we calculated the value function of the corresponding estimated optimal rule using expression (1) as the empirical average of a large independent test data set with a sample size of 20 000.

3.2 | Simulation results

The results from 500 replicates are presented in Figures 1 and 2 and Table 1. In both simulation settings and for all sample sizes, AOL shows a significant advantage over OWL in terms of a higher value function because of augmentation and other improvements highlighted in previous sections. In the first setting, we observe that Q-learning has a higher value function than AOL but also a higher variability with a smaller sample size ($n = 50$, $n =$ 100, and $n = 200$). With a large sample size ($n = 400$), Q-learning has an empirical standard deviation smaller than AOL. The value function for both Q-learning and AOL increases with the sample size, where the former increases at a faster rate. In this setting, the linear regression model is a good approximation for the Q-function since the rewards were generated from a linear model. Therefore, Q-learning may achieve the theoretical optimal value faster than AOL when n increases. Comparing with OWL, AOL achieves a much larger value and a smaller standard deviation for all sample sizes. In the second simulation setting, the optimal treatment boundaries were more complicated and highly nonlinear; therefore, Q-learning performed the worst among the three method at all sample sizes. For example, it only achieves a median value of 0.717 when $n = 400$ compared with the true optimal value of 4. For a proportion of the 500 replications, no treatment by covariate interaction terms were selected by Lasso regression in at least one step of Q-learning. In this case, the optimal treatment was selected randomly to compute the value function using the test data. Moreover, AOL outperforms OWL and Q-learning in all cases and achieves a median value of 3.211 with a sample size of 400.

For the previous two simulation scenarios, we also implemented AOL with Gaussian kernel with four-fold cross validation to choose the bandwidth and compared with OWL with Gaussian kernel. Moreover, AOL with Gaussian kernel performed similarly with linear kernel for AOL in the second scenario, which achieved a median of 0.851, 1.453, 2.427, 3.400 and for the sample size of 50, 100, 200, and 400, respectively. For the first scenario, AOL with Gaussian kernel has a slightly worse performance than linear kernel, with a median value function of 4.965, 5.496, 6.295, 6.905 for sample sizes ranging from 50 to 400. Comparatively, OWL with Gaussian kernel has the median values of 2.529, 2.952, 3.459, 4.288 in the first scenario and 0.804, 1.043, 1.356, 1.743 in the second scenario, so AOL is still superior to OWL. Since the computational burden for the Gaussian kernel is heavier, we conclude that using a linear kernel for AOL is sufficient in these simulation settings. Additional simulations of the single-stage and two-stage settings are reported in the supplementary material (Section B). Similar comparative performances are observed.

In summary, in simulation scenario 1, the data were generated such that a linear function is an adequate approximation for the true cumulative rewards. Thus, Q-learning outperformed OWL and AOL. In simulation scenario 2, the data were generated such that the cumulative rewards cannot be approximated adequately by the linear models. Thus, OWL and AOL outperformed the value-based learning method Q-learning. Nevertheless, in all the presented simulation scenarios, AOL outperformed OWL, which demonstrates that the proposed AOL improves OWL.

4 | REAL DATA APPLICATION

We applied the proposed method to data from the STAR*D trial,[4] which was a phase-IV multisite, prospective, multistage, randomized clinical trial to compare various treatment regimes for patients with nonpsychotic MDD.[4] The detail of the study design is given in the supplementary. The aim of STAR∗D was to find the best subsequent treatment for subjects who failed to achieve adequate response to an initial antidepressant treatment (citalopram). The primary outcome was measured by the Quick Inventory of Depressive Symptomatology (QIDS) score ranging from 0 to 26 in the sample. Participants with a total clinician-rated QIDS score under 5 were considered as having a clinically meaningful response to the treatment and therefore in remission. Remitted patients were not eligible for any future treatments and entered a follow-up phase.

Following the works of Chakraborty and Moodie[14] and Pineau et al,[23] we focused on a two-stage decision-making problem by combining study levels 2 and 2A as the first stage and treating study level 3 as the second stage. Additionally, different drugs were combined as one class of drugs involving selective serotonin reuptake inhibitors (SSRI) and the other class of drugs without SSRI. Thus, at each stage, treatment (A_k), reward outcome (R_k), and feature variables (H_k) were defined as follows:

A_1: 1 if SSRI drugs are used and −1 SSRI drugs are not used at level 2 and 2A (stage 1);

A_2: 1 if SSRI drugs are used and −1 SSRI drugs are not used at at level 3 (stage 2);

R_1: -QIDS score at the end of first stage if remission was achieved, −½ QIDS score at the end of first stage if remission was not achieved;

R_2: −½ QIDS score at the end of second stage;

H_1: baseline QIDS score (at the beginning of the trial), the rate of change of QIDS score from baseline to stage 1 randomization (level 1 to level 2), participant preference (taking values −1, 0, or 1), and QIDS at the beginning of stage 1 randomization;

H_2: H_1, the rate of change of QIDS score during stage 1, participant preference at stage 2 randomization, A_1, and its interactions with the previous variables.

There were 1381 participants with complete feature variables for the first stage analysis, among whom 516 achieved remission at the end of the first stage. Among 865 nonremitted participants, 364 of them had entered the second stage and have complete information on the feature variables and outcomes. In the analysis, the patients who had remission in stage 1 were treated as if they would have received the optimal treatments at stage 2; thus, we only analyzed patients who had entered stage 2 in order to estimate the optimal rule at stage 2.

To implement AOL, we followed the steps in Section 2.3, where the first-stage randomization probability π_1 was calculated as the frequency of SSRI and non-SSRI given patient preference at stage 1 and the second stage randomization probability π_2 was computed as the similar frequency and further multiplied by the nondropout proportions to account for missingness in this stage. More specifically, Lasso regression was implemented in Step 2–1, and a weighted Lasso was used for Step 2–1. For comparison, we also implemented Q-learning and OWL, where Lasso was used in the regression in each stage of Q-learning; the same π_1 and π_2 were used. Both gaussian and linear kernels were implemented for AOL and OWL. Comparison of all the methods were based on 1000 repetitions of two-fold cross-validation: for each cross-validation, one-half data were used for training, and the other half were used to compute the value functions for the estimated DTRs. For each replication, the testing value function was computed as the empirical estimation following Equation (1), which is the weighted average of the cumulative rewards for all patients whose observed treatments agree with the estimated optimal treatments in all stages.

Q-learning, OWL, and AOL were compared in Figure 3. The mean baseline clinician-rated QIDS score in the sample was 16.71, and the mean QIDS at the start of stage 1 randomization was 12.37. The average testing QIDS score for the optimal DTR obtained by AOL with Gaussian kernel was 6.733 points (sd=4.08), which outperformed Q-learning (7.93, sd=2.38) and OWL with Gaussian kernel (10.85, sd=1.11). Gaussian kernel yielded better testing value than linear kernel for both AOL and OWL; AOL with linear kernel had an average testing value of 8.38 (sd=3.10), which was still better than OWL with linear kernel (10.85, sd=0.99). Moreover, AOL-estimated rule also outperformed the one-size-fits-all rules (eg, all subjects receive SSRI in both stages, all subjects received SSRI in the first stage and non-SSRI in the second and so on).

Furthermore, we examine the coefficients of AOL fitted by a linear kernel using the standardized feature variables. We present the normalized effects for the optimal DTR obtained by AOL in Figure 4. We normalized the effect of each tailoring variable through dividing by the L_2 norm of all coefficients of the decision rule. The baseline variables at first stage with strongest effects were baseline QIDS score, rate of change of QIDS in the previous period, and patient preference. The strongest second-stage tailoring variables were intermediate outcome after stage 1 treatment, starting QIDS at stage 1, and patient preference for the second-stage treatment.

In conclusion, the STAR∗D example demonstrates that AOL outperforms the alternative methods in maximizing the clinical benefits, and it also yields some insights on combining tailoring variables for deep tailoring and forming new treatment rules.

5 | DISCUSSION

In this work, we propose a new machine learning method, AOL, to estimate optimal DTRs through robust and efficient augmentation to OWL. We theoretically prove that AOL guarantees efficiency improvement over OWL for both $K = 1$ and $K > 1$ stages. The theoretical results show that AOL has the same approximation bias but a smaller stochastic error. Moreover, AOL achieves efficiency gain by properly constructing surrogate outcomes with smaller second moments. In an earlier version of this paper (https://arxiv.org/abs/1611.02314), we provided an additional application to a SMART of children affected by attention deficit and hyperactive disorder. A recent publication by Zhang and Zhang[24] considered similar augmented outcomes as weights in multiple-stage estimation but used genetic algorithm for estimating optimal treatments. In comparison, our proposed method used computationally more stable large margin loss, and we rigorously justified the advantage of the proposed method in terms of the risk bound for the value function.

In real-world studies, it may be difficult to identify a priori which variables may serve as tailoring variables for treatment response. In our simulation studies, AOL has shown to be superior in such settings with non-treatment-differentiating noise variables and unknown treatment mechanisms. In addition, using a more sophisticated prediction method (eg, random forest) to incorporate nonlinear interactions between health history variables H_k to predict $s_k(H_k)$ in the step of taking residuals may be beneficial, although theoretically, a linear model will guarantee improved efficiency of AOL over OWL.

Clinicians may be interested in ranking the most important variables to predict patient heterogeneity to treatment. Biomarkers that could signal patients' heterogeneous responses to various interventions are especially useful as tailoring variables. This information can be used to design new intervention arms in future confirmatory trials and facilitate discovering new knowledge in medical research. Variable selection may help construct a less noisy rule and avoid over-fitting. Although AOL leads to a sparse DTR in the STAR∗D example, a future research topic is to investigate methods that perform automatic variable selection in the outcome-weighted learning framework. Additionally, our current framework can easily handle nonlinear decision functions by using nonlinear kernels, which may improve performance for high-dimensional correlated tailoring variables. It is also of interest to consider other kinds of decision functions such as decision trees to construct DTRs that are highly interpretable.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding information

National Institute of Health, Grant/Award Number: 1R01GM124104-01A1, 2R01NS073671-05A1, 5P01CA142538-07, 1R01DK108073-01

REFERENCES

1. Lavori PW, Dawson R. A design for testing clinical strategies: biased adaptive within-subject randomization. J Royal Stat Soc Ser A Stat Soc. 2000;163(1):29–38.
2. Thall PF, Sung HG, Estey EH. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. J Am Stat Assoc. 2002;97(457):29–39.
3. Lunceford JK, Davidian M, Tsiatis AA. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. Biometrics. 2002;58(1):48–57. [PubMed: 11890326]
4. Rush AJ, Fava M, Wisniewski SR, et al. Sequenced treatment alternatives to relieve depression (STAR* D): rationale and design. Contemp Clin Trials. 2004;25(1):119–142.
5. Murphy SA. An experimental design for the development of adaptive treatment strategies. Stat Med. 2005;24(10):1455–1481. [PubMed: 15586395]
6. Moodie EEM, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. Biometrics. 2007;63(2):447–455. [PubMed: 17688497]
7. Murphy SA. Optimal dynamic treatment regimes. J Royal Stat Soc Ser B Stat Methodol. 2003;65(2):331–355.
8. Robins JM. Optimal structural nested models for optimal sequential decisions In: Proceedings of the Second Seattle Symposium in Biostatistics. New York, NY: Springer; 2004:189–326.
9. Zhao Y, Kosorok MR, Zeng D. Reinforcement learning design for cancer clinical trials. Stat Med. 2009;28(26):3294–3315. [PubMed: 19750510]
10. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. J Am Stat Assoc. 2012;107(499):1106–1118. [PubMed: 23630406]
11. Wang L, Rotnitzky A, Lin X, Millikan RE, Thall PF. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. J Am Stat Assoc. 2012;107(498):493–508. [PubMed: 22956855]
12. Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Ann Stat. 2011;39(2):1180–1210. [PubMed: 21666835]
13. Kang C, Janes H, Huang Y. Combining biomarkers to optimize patient treatment recommendations. Biometrics. 2014;70(3):695–707. [PubMed: 24889663]
14. Chakraborty B, Moodie EEM. Statistical Methods for Dynamic Treatment Regimes. New York, NY: Springer; 2013.
15. Kosorok MR, Moodie EM. Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine. Philadelphia, PA: SIAM; 2016.
16. Watkins CJCH. Learning from Delayed Rewards [PhD thesis]. Cambridge, UK: University of Cambridge; 1989.
17. Murphy SA, Oslin DW, Rush AJ, Zhu J. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. Neuropsychopharmacology. 2006;32(2): 257–262. [PubMed: 17091129]

18. Zhao Y, Zeng D, Laber E, Kosorok MR. New statistical learning methods for estimating optimal dynamic treatment regimes. J Am Stat Assoc. 2015;110(510):583–598. 10.1080/01621459.2014.937488 [PubMed: 26236062]

19. Murphy SA, van der Laan MJ, Robins JM. Marginal mean models for dynamic regimes. J Am Stat Assoc. 2001;96(456):1410–1423. [PubMed: 20019887]

20. Zeileis A, Hornik K, Smola A, Karatzoglou A. kernlab-an S4 package for kernel methods in R. J Stat Softw. 2004;11(9):1–20.

21. Tsiatis AA. Semiparametric Theory and Missing Data. New York, NY: Springer; 2006.

22. Zhang B, Tsiatis AA, Laber EB, Davidian M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. Biometrika. 2013;100(3):681–694.

23. Pineau J, Bellemare MG, Rush AJ, Ghizaru A, Murphy SA. Constructing evidence-based treatment strategies using methods from computer science. Drug Alcohol Depend. 2007;88:S52–S60. [PubMed: 17320311]

24. Zhang B, Zhang M. C-learning: a new classification framework to estimate optimal dynamic treatment regimes. Biometrics. 2018 In press.

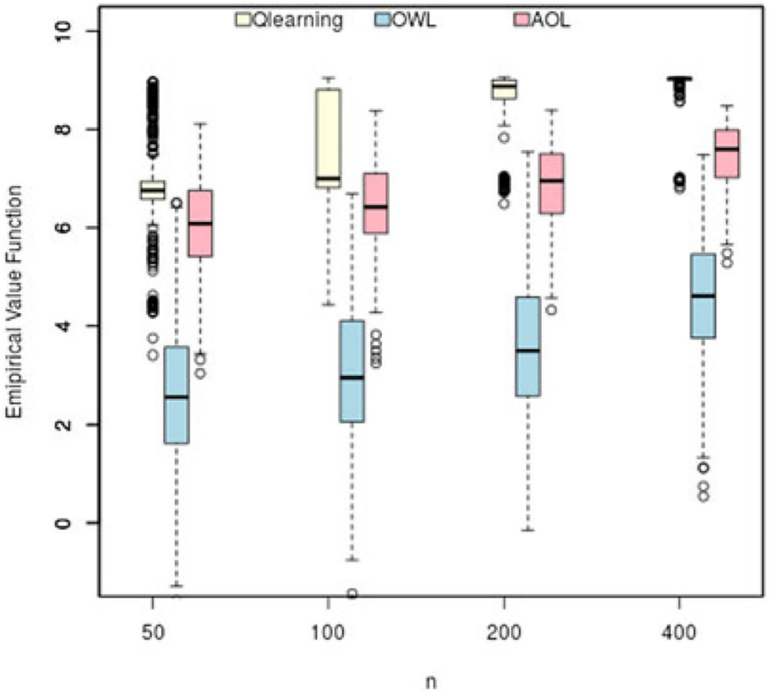

FIGURE 1.

Simulation setting 1 with four-stage design (optimal value = 10.1). AOL, Augmented Outcome-weighted Learning. OWL, outcome-weighted learning [Colour figure can be viewed at wileyonlinelibrary.com]

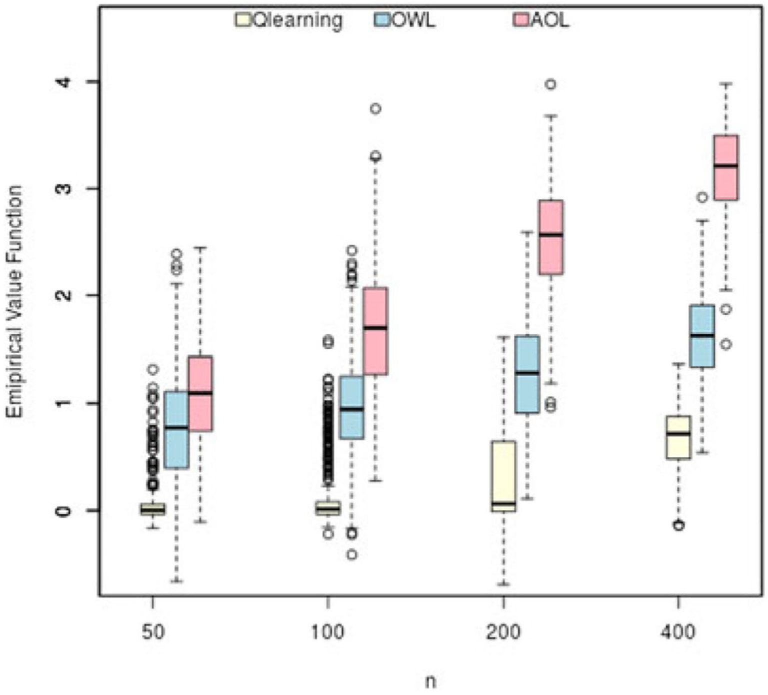

FIGURE 2.

Simulation setting 2 with four-stage design (optimal value = 4.0). AOL, Augmented Outcome-weighted Learning; OWL, outcome-weighted learning [Colour figure can be viewed at wileyonlinelibrary.com]

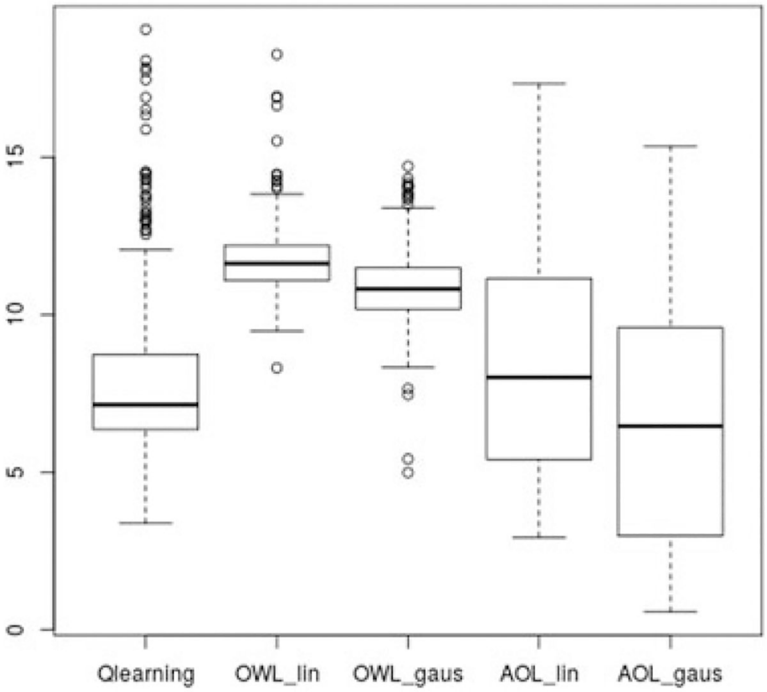

FIGURE 3.

Mean and standard error of the value function (depression symptom score, Quick Inventory of Depressive Symptomatology) based on 1000 repetitions of two-fold cross validation for Sequenced Treatment Alternatives to Relieve Depression data (lower score desirable)

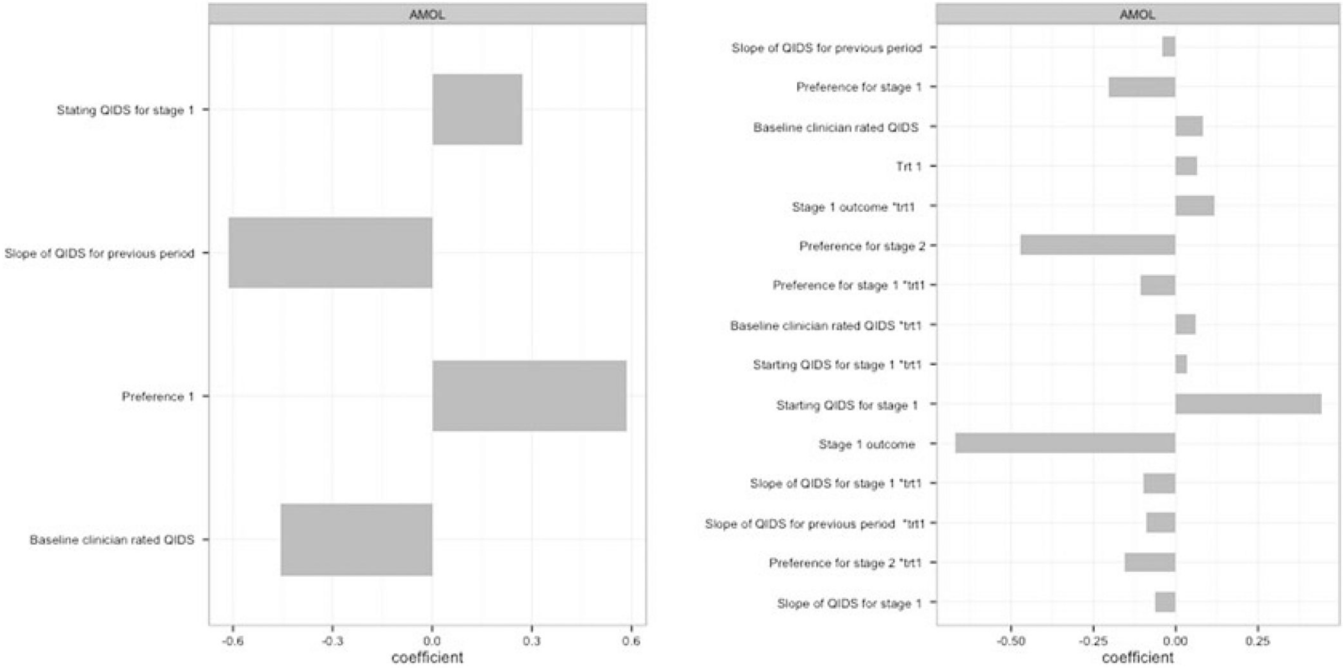

FIGURE 4.

Normalized coefficients of the stage 1 tailoring variables (left panel) and stage 2 tailoring variables (right panel) obtained by Augmented Outcome-weighted Learning. QIDS, Quick Inventory of Depressive Symptomatology

TABLE 1

Mean and median of the empirical value function for two simulation scenarios evaluated with an independent test data set

Simulation Setting 1 (Optimal Value 10.1)						
n	**Q-learning**		**OWL**		**AOL**	
	Mean (SD)	**Median**	**Mean (SD)**	**Median**	**Mean (SD)**	**Median**
50	6.786(1.119)	6.753	2.604(1.502)	2.561	6.042(0.951)	6.078
100	7.711(1.016)	6.996	3.049(1.448)	2.957	6.436(0.859)	6.415
200	8.475(0.843)	8.874	3.593(1.461)	3.486	6.865(0.756)	6.949
400	8.934(0.398)	9.034	4.566(1.265)	4.603	7.467(0.632)	7.593

Simulation Setting 2 (Optimal Value 4)						
n	**Q-learning**		**OWL**		**AOL**	
	Mean (SD)	**Median**	**Mean (SD)**	**Median**	**Mean (SD)**	**Median**
50	0.042(0.182)	0.003	0.764(0.522)	0.773	1.105(0.522)	1.097
100	0.103(0.281)	0.011	0.966(0.484)	0.944	1.696(0.595)	1.698
200	0.291(0.404)	0.062	1.281(0.492)	1.284	2.519(0.518)	2.568
400	0.635(0.355)	0.717	1.638(0.446)	1.626	3.177(0.421)	3.211

Abbreviations: AOL, Augmented Outcome-weighted Learning; OWL, outcome-weighted learning.