Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Interactive Q-learning for Quantiles

Kristin A. Linn, Eric B. Laber & Leonard A. Stefanski

To cite this article: Kristin A. Linn, Eric B. Laber & Leonard A. Stefanski (2016): Interactive Q-learning for Quantiles, Journal of the American Statistical Association, DOI: 10.1080/01621459.2016.1155993

To link to this article: http://dx.doi.org/10.1080/01621459.2016.1155993

Interactive Q-learning for Quantiles

Kristin A. Linn[1], Eric B. Laber[2], Leonard A. Stefanski[2]
[1]Department of Biostatistics and Epidemiology
University of Pennsylvania, Philadelphia, PA 19104
[2]Department of Statistics
North Carolina State University, Raleigh, NC 27695
email: klinn@upenn.edu

Author's Footnote:

Kristin A. Linn is Postdoctoral Fellow, Department of Biostatistics and Epidemiology, University of Pennsylvania. Mailing address: University of Pennsylvania School of Medicine, Center for Clinical Epidemiology and Biostatistics (CCEB), Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104 (email: klinn@upenn.edu); Eric B. Laber is Assistant Professor, Department of Statistics, North Carolina State University; Leonard A. Stefanski is Drexel Professor, Department of Statistics, North Carolina State University. Eric Laber acknowledges support from NIH grant P01 CA142538 and DNR grant PR-W-F14AF00171. Leonard Stefanski acknowledges support from NIH grants R01 CA085848 and P01 CA142538 and NSF grant DMS-0906421.

Abstract

A dynamic treatment regime is a sequence of decision rules, each of which recommends treatment based on features of patient medical history such as past treatments and outcomes. Existing methods for estimating optimal dynamic treatment regimes from data optimize the mean of a response variable. However, the mean may not always be the most appropriate summary of performance. We derive estimators of decision rules for optimizing probabilities and quantiles computed with respect to the response distribution for two-stage, binary treatment settings. This enables estimation of dynamic treatment regimes that optimize the cumulative distribution function of the response at a prespecified point or a prespecified quantile of the response distribution such as the median. The proposed methods perform favorably in simulation experiments. We illustrate our approach with data from a sequentially randomized trial where the primary outcome is remission of depression symptoms.

KEYWORDS: Dynamic Treatment Regime; Personalized Medicine; Sequential Decision Making; Sequential Multiple Assignment Randomized Trial.

1. INTRODUCTION

A dynamic treatment regime operationalizes clinical decision making as a series of decision rules that dictate treatment over time. These rules account for accrued patient medical history, including past treatments and outcomes. Each rule maps current patient characteristics to a recommended treatment, hence personalizing treatment. Typically, a dynamic treatment regime is estimated from data with the goal of optimizing the expected value of a clinical outcome, and the resulting regime is referred to as the estimated optimal regime.

Direct-search, also known as policy-search or value-search, is one approach to estimating an optimal dynamic treatment regime. Direct search estimators require a pre-specified class of dynamic treatment regimes and an estimator of the marginal mean outcome under any regime in the pre-specified class. The maximizer of the estimated marginal mean outcome over the class of regimes is taken as the estimator of the optimal dynamic treatment regime. Marginal structural models (MSMs) are one type of direct-search estimators (Robins, 2000; van der Laan et al., 2005; van der Laan, 2006; van der Laan and Petersen, 2007; Bembom and van der Laan, 2008; Robins et al., 2008; Orellana et al., 2010; Petersen et al., 2014). MSMs are best-suited to problems with a small class of potential regimes. MSMs may also be advantageous in practice because optimizing over a small class of pre-specified regimes provides a simpler, and often more interpretable, regime than other approaches. Another class of direct-search estimators casts the marginal mean outcome as a weighted missclassification rate and applies either discrete-optimization or classification algorithms to optimize a plugin estimator of the marginal mean outcome (Zhao et al., 2012; Zhang et al., 2012b,a, 2013; Zhao et al., 2015).

Regression-based or indirect estimators comprise a second class of estimators of an optimal dynamic treatment regime. Regression-based estimators require a model for some portion of the conditional distribution of the outcome given treatments and covariate information. Examples of regression-based estimators include Q-learning (Watkins, 1989; Watkins and Dayan, 1992; Murphy, 2005a), regularized Q-learning (Moodie and Richardson, 2010; Chakraborty et al., 2010;

Song et al., 2015; Goldberg et al., 2013), Interactive Q-learning (Laber et al., 2014a), g-estimation in structural nested mean models (Robins, 2004), A-learning (Murphy, 2003), and regret-regression (Henderson et al., 2010). Nonparametric regression-based approaches often target the globally optimal regime rather than restricting attention to a small, pre-specified class. They can be also be useful in exploratory contexts to discover new treatment strategies for further evaluation in later trials.

Direct-search and regression-based estimators have been extended to handle survival outcomes (Goldberg and Kosorok, 2012; Huang and Ning, 2012; Huang et al., 2014), high-dimensional data (McKeague and Qian, 2013), missing data (Shortreed et al., 2014), multiple outcomes (Laber et al., 2014b; Linn et al., 2015), and restrictions on treatment resource (Luedtke and van der Laan, 2015).

Despite many estimation methods, none are designed to handle functionals of the response distribution other than the mean, such as quantiles. The median response is often of interest in studies where the outcome follows a skewed distribution, such as the total time a women spends in second stage labor (Zhang et al., 2012c). Using the potential outcomes framework (Rubin, 1974; Rosenbaum and Rubin, 1983), Zhang et al. (2012c) develop methods for estimating quantiles of the potential outcomes from observational data. However, they focus on comparing treatments at a single intervention time point rather than estimation of an optimal dynamic treatment regime. Structural nested distribution models (SNDMs) estimated using g-estimation facilitate estimation of point treatment effects on the cumulative distribution function of the outcome (Robins, 2000; Vansteelandt et al., 2014). Thus far, SNDMs have not been extended to estimate a regime that maximizes a quantile.

Q-learning and its variants are often useful when targeting an optimal regime because they provide relatively interpretable decision rules that are based on (typically linear) regression models. However, the Q-learning algorithm is an approximate dynamic programming procedure that requires modeling nonsmooth, nonmonotone transformations of data. This leads to nonregular estimators for parameters that index the optimal regime and complicates the search for models that

fit the data well since many standard regression modeling diagnostics are invalid (Robins, 2004; Chakraborty et al., 2010; Laber et al., 2014c; Song et al., 2015). In addition, Q-learning with linear models does not target the globally optimal rule when the true conditional means are nonlinear. Interactive Q-learning (IQ-learning), developed for the two-stage binary treatment setting, requires modeling only smooth, monotone transformations of the data, thereby reducing problems of model misspecification and nonregular inference (Laber et al., 2014a). We extend the IQ-learning framework to optimize functionals of the outcome distribution other than the expected value. In particular, we optimize threshold-exceedance probabilities and quantiles of the response distribution. Furthermore, because this extension of IQ-learning provides an estimator of a threshold-exceedance probability or quantile of the response distribution under any postulated dynamic treatment regime, it can be used to construct direct-search estimators.

Threshold-exceedance probabilities are relevant in clinical applications where the primary objective is remission or a specific target for symptom reduction. For example, consider a population of obese patients enrolled in a study to determine the effects of several treatment options for weight loss. The treatments of interest may include combinations of drugs, exercise programs, counseling, and meal plans (Berkowitz et al., 2010). Our method can be used to maximize the probability that patients achieve a weight below some prespecified, patient-specific threshold at the conclusion of the study. Optimization of threshold-exceedance probabilities can be framed as a special case optimizing the mean of a binary outcome, for which several methods exist, including the classification-based Outcome Weighted Learning (Zhang et al., 2012b; Zhao et al., 2015). However, our approach is particularly useful for setting up the more challenging problem of quantile optimization.

With adjustments to our method of maximizing probabilities, we derive optimal decision rules for maximizing quantiles of the response distribution. Both frameworks can be used to study the *entire distribution* of the outcome under an optimal dynamic treatment regime; thus, investigators can examine how the optimal regime changes as the target probability or quantile is varied. In

addition, the quantile framework provides an analog of quantile regression in the dynamic treatment regime setting for constructing robust estimators; for example, it enables optimization of the median response.

2. GENERALIZED INTERACTIVE Q-LEARNING

We first characterize the optimal regime for a probability and quantile using potential outcomes (Rubin, 1974) and two treatment time-points. We assume that the observed data, $\mathcal{D} = \{(\boldsymbol{X}_{1i}, A_{1i}, \boldsymbol{X}_{2i}, A_{2i}, Y_i)\}_{i=1}^n$, comprise n independent, identically distributed, time-ordered trajectories; one per patient. Let $(\boldsymbol{X}_1, A_1, \boldsymbol{X}_2, A_2, Y)$ denote a generic observation where: $\boldsymbol{X}_1 \in \mathbb{R}^{p_1}$ is baseline covariate information collected prior to the first treatment; $A_1 \in \{-1, 1\}$ is the first treatment; $\boldsymbol{X}_2 \in \mathbb{R}^{p_2}$ is interim covariate information collected during the course of the first treatment but prior the second treatment; $A_2 \in \{-1, 1\}$ is the second treatment; and $Y \in \mathbb{R}$ is an outcome measured at the conclusion of stage two, coded so that larger is better. Define $\boldsymbol{H}_1 = \boldsymbol{X}_1$ and $\boldsymbol{H}_2 = (\boldsymbol{H}_1^\intercal, A_1, \boldsymbol{X}_2^\intercal)^\intercal$ so that $\boldsymbol{H}_t$ is the information available to a decision maker at time t. A regime, $\boldsymbol{\pi} = (\pi_1, \pi_2)$, is a pair of decision rules where $\pi_t : \text{dom}(\boldsymbol{H}_t) \mapsto \text{dom}(A_t)$, such that a patient presenting with $\boldsymbol{H}_t = \boldsymbol{h}_t$ at time t is recommended treatment $\pi_t(\boldsymbol{h}_t)$.

Let $\boldsymbol{H}_2^*(a_1)$ be the potential second-stage history under treatment a_1 and $Y^*(a_1, a_2)$ the potential outcome under treatment sequence (a_1, a_2). Define the set of all potential outcomes $W = \left\{\boldsymbol{H}_2^*(a_1), Y^*(a_1, a_2) : (a_1, a_2) \in \{-1, 1\}^2\right\}$. Throughout we assume: (C1) consistency, so that $Y = Y^*(A_1, A_2)$; (C2) sequential ignorability (Robins, 2004), i.e., $A_t \perp\!\!\!\perp W \mid \boldsymbol{H}_t$ for $t = 1, 2$; and (C3) positivity, so that there exists $\epsilon > 0$ for which $\epsilon < \text{pr}(A_t = a_t|\boldsymbol{H}_t) < 1 - \epsilon$ with probability one for all a_t, $t = 1, 2$. Assumptions (C2)-(C3) hold by design when data are collected using a sequential multiple assignment randomized trial (SMART, Lavori and Dawson, 2000, 2004; Murphy, 2005b). In observational studies, these assumptions are not testable. We assume that data are collected using a two-stage, binary treatment SMART. This set-up facilitates a focused discussion of the proposed methods and is also useful in practice, as data in many sequentially randomized trials have this

structure (Projects Using SMART, 2012; Laber, 2013). However, the following argument demonstrates that the proposed methodology can be extended to observational data and studies with more than two treatments.

For any $\boldsymbol{\pi}$ define $Y^*(\boldsymbol{\pi}) = \sum_{(a_1,a_2)} Y^*(a_1,a_2)\mathbb{1}_{\pi_1(\boldsymbol{H}_1)=a_1}\mathbb{1}_{\pi_2\{\boldsymbol{H}_2^*(a_1)\}=a_2}$ to be the potential outcome under $\boldsymbol{\pi}$. Define the function $R(y;\boldsymbol{x}_1,a_1,\boldsymbol{x}_2,a_2) = \text{pr}(Y > y|\boldsymbol{X}_1 = \boldsymbol{x}_1, A_1 = a_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, A_2 = a_2)$ for any $y \in \mathbb{R}$. Assuming (C1)-(C3) and $y \in \mathbb{R}$, the survival function of $Y^*(\boldsymbol{\pi})$ can be expressed in terms of the underlying generative model as

$$\text{pr}\{Y^*(\boldsymbol{\pi}) > y\} = \mathbb{E}\left[\sum_{a_1}\mathbb{1}_{\pi_1(\boldsymbol{H}_1)=a_1}\mathbb{E}\left\{\sum_{a_2}\mathbb{1}_{\pi_2(\boldsymbol{H}_2)=a_2}R(y;\boldsymbol{X}_1,a_1,\boldsymbol{X}_2,a_2)\Big|\boldsymbol{X}_1,A_1=a_1\right\}\right]$$

for any $\boldsymbol{\pi}$ (Robins, 1986). This result shows that $\text{pr}\{Y^*(\boldsymbol{\pi}) > y\}$ is maximized by the regime $\boldsymbol{\pi}^y = (\pi_1^y, \pi_2^y)$, where $\pi_2^y(\boldsymbol{h}_2) = \arg\max_{a_2} R(y;\boldsymbol{x}_1,a_1,\boldsymbol{x}_2,a_2)$ and $\pi_1^y(\boldsymbol{h}_1) = \arg\max_{a_1}\mathbb{E}\left\{\sum_{a_2}\mathbb{1}_{\pi_2^y(\boldsymbol{H}_2)=a_2}R(y;\boldsymbol{x}_1,a_1,\boldsymbol{X}_2,a_2) \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, A_1 = a_1\right\}$. This result can also be used to characterize the regime that optimizes a quantile. For any regime $\boldsymbol{\pi}$, $\text{pr}\{Y^*(\boldsymbol{\pi}^y) \le y\} \le \text{pr}\{Y^*(\boldsymbol{\pi}) \le y\}$ implies $\inf\{y : \text{pr}\{Y^*(\boldsymbol{\pi}^y) \le y\} \ge \tau\} \ge \inf\{y : \text{pr}\{Y^*(\boldsymbol{\pi}) \le y\} \ge \tau\}$. In fact the left-hand side is the optimal τ^{th} quantile; denote it by y_τ^*. An optimal regime with respect to the τ^{th} quantile is thus any regime that, when used to assign treatments in the population, results in a τ^{th} quantile that attains y_τ^*. The following result is proved in the supplemental material.

Theorem 2.1. *Let $\epsilon > 0$ and $\tau \in (0,1)$ be arbitrary but fixed. Assume (C1)-(C3) and that the map $y \mapsto R(y;\boldsymbol{x}_1,a_1,\boldsymbol{x}_2,a_2)$ from $\mathbb{R}$ into $(0,1)$ is continuous and strictly increasing in a neighborhood of τ for all $\boldsymbol{x}_1, a_1, \boldsymbol{x}_2$, and a_2. Then,* $\inf\{y : pr\left\{Y^*(\boldsymbol{\pi}^{y_\tau^*}) \le y\right\} \ge \tau\} = y_\tau^*$.

Theorem 2.1 states that the regime $\boldsymbol{\pi}^{y_\tau^*}$ induces a τ^{th} quantile of the potential outcome distribution that attains the optimal τ^{th} quantile. Without the strictly increasing assumption, $\boldsymbol{\pi}^{y_\tau^*}$ may not be optimal. However, it can be shown that there exists a value $\tilde{y} \in \mathbb{R}$ such that the regime $\boldsymbol{\pi}^{\tilde{y}}$ attains a τ^{th} quantile that is arbitrarily close to y_τ^*. Details are given in Section 3.

2.1 Threshold Interactive Q-learning

The estimators presented in this section serve as useful building blocks for developing the estimators in the next section which focus on quantile optimization. Here, we derive and estimate the optimal set of decision rules for maximizing a threshold-exceedance probability. Let $\text{pr}^{\boldsymbol{\pi}}(Y > \lambda)$, equivalently $\text{pr}^{\pi_1,\pi_2}(Y > \lambda)$, denote the probability that the outcome Y is greater than a predefined threshold λ under treatment assignment dictated by the regime $\boldsymbol{\pi} = (\pi_1, \pi_2)$. Threshold Interactive Q-learning (TIQ-learning) maximizes $\text{pr}^{\boldsymbol{\pi}}\{Y > \lambda(\boldsymbol{H}_t) \mid \boldsymbol{H}_1\}$ for all $\boldsymbol{H}_1$ with respect to $\boldsymbol{\pi}$, where $\lambda(\boldsymbol{H}_t)$ is a threshold that depends on $\boldsymbol{H}_t$, $t = 1, 2$. Here, we assume $\lambda(\boldsymbol{H}_t) \equiv \lambda$; patient-specific thresholds are discussed in the supplemental material.

As $\text{pr}^{\boldsymbol{\pi}}(Y > \lambda) = E^{\boldsymbol{\pi}}(\mathbb{1}_{Y>\lambda})$, many approaches exist for estimating an optimal regime that maximizes the value function of the binary outcome $\mathbb{1}_{Y>\lambda}$; one example is discrete Q-learning (Chakraborty and Moodie, 2013). We show analytically in Remark 2.4, and empirically in Section 4, that discrete Q-learning using the logit link is equivalent to Q-learning with outcome Y and is therefore insensitive to the threshold λ. Nonparametric methods such as OWL (Zhao et al., 2012) double-robust direct search (Zhang et al., 2012b,a, 2013), and boosting (Kang et al., 2014) are less restrictive than model-based approaches for optimizing binary outcomes, but the following exposition facilitates later developments for optimizing quantiles.

Our estimators are derived under the following set-up. Because A_2 is binary, there exist functions m and c such that $E(Y \mid A_2, \boldsymbol{H}_2) = m(\boldsymbol{H}_2) + A_2 c(\boldsymbol{H}_2)$. We assume that $Y = E(Y \mid A_2, \boldsymbol{H}_2) + \epsilon$, where $E(\epsilon) = 0$, $\text{Var}(\epsilon) = \sigma^2$, and ϵ is independent of $(A_2, \boldsymbol{H}_2)$. In the supplemental material we describe extensions to: (i) heteroskedastic error structures, where $Y = E(Y \mid A_2, \boldsymbol{H}_2) + \sigma(\boldsymbol{H}_2, A_2)\epsilon$ for unknown function σ; and (ii) non-additive error structures such as the multiplicative error model, $Y = \epsilon[m(\boldsymbol{H}_2) + A_2 c(\boldsymbol{H}_2)]$, provided $\epsilon > 0$ with probability one and $\text{pr}\{m(\boldsymbol{H}_2) + A_2 c(\boldsymbol{H}_2) = 0\} = 0$.

Define $F_{\boldsymbol{H}_1}(\cdot)$ to be the distribution of $\boldsymbol{H}_1$; $F_{\boldsymbol{H}_2 \mid \boldsymbol{H}_1, A_1}(\cdot \mid \boldsymbol{h}_1, a_1)$ to be the conditional distribution of $\boldsymbol{H}_2$ given $\boldsymbol{H}_1 = \boldsymbol{h}_1$ and $A_1 = a_1$; $F_\epsilon(\cdot)$ to be the distribution of ϵ; and $\boldsymbol{H}_2^{\pi_1(\boldsymbol{H}_1)} =$

$\{\boldsymbol{H}_1^\intercal, \pi_1(\boldsymbol{H}_1), \boldsymbol{X}_2^\intercal\}^\intercal$. Let $J^{\pi_1,\pi_2}(\boldsymbol{h}_1, \boldsymbol{h}_2, y) = F_\epsilon\{y - m(\boldsymbol{h}_2^{\pi_1(\boldsymbol{h}_1)}) - \pi_2(\boldsymbol{h}_2^{\pi_1(\boldsymbol{h}_1)})c(\boldsymbol{h}_2^{\pi_1(\boldsymbol{h}_1)})\}$, then

$$\text{pr}^{\pi_1,\pi_2}(Y \leq y) = \int\int J^{\pi_1,\pi_2}(\boldsymbol{h}_1, \boldsymbol{h}_2, y) dF_{\boldsymbol{H}_2 \mid \boldsymbol{H}_1, A_1}\{\boldsymbol{h}_2 \mid \boldsymbol{h}_1, \pi_1(\boldsymbol{h}_1)\} dF_{\boldsymbol{H}_1}(\boldsymbol{h}_1), \tag{1}$$

is the expected value of $J^{\pi_1,\pi_2}(\boldsymbol{H}_1, \boldsymbol{H}_2, y)$.

Let $\pi_2^*(\boldsymbol{h}_2) = \text{sgn}\{c(\boldsymbol{h}_2)\}$, where $\text{sgn}(x) = \mathbb{1}_{x\geq 0} - \mathbb{1}_{x<0}$. Then, $J^{\pi_1,\pi_2^*}(\boldsymbol{h}_1, \boldsymbol{h}_2, y) = F_\epsilon\{y - m(\boldsymbol{h}_2^{\pi_1(\boldsymbol{h}_1)}) - |c(\boldsymbol{h}_2^{\pi_1(\boldsymbol{h}_1)})|\}$ and $\pi_2(\boldsymbol{h}_2^{\pi_1(\boldsymbol{h}_1)})c(\boldsymbol{h}_2^{\pi_1(\boldsymbol{h}_1)}) \leq |c(\boldsymbol{h}_2^{\pi_1(\boldsymbol{h}_1)})|$ for all $\boldsymbol{h}_2^{\pi_1(\boldsymbol{h}_1)}$, implies

$$\text{pr}^{\pi_1,\pi_2}(Y \leq y) \;\geq\; \int\int J^{\pi_1,\pi_2^*}(\boldsymbol{h}_1, \boldsymbol{h}_2, y) dF_{\boldsymbol{H}_2 \mid \boldsymbol{H}_1, A_1}\{\boldsymbol{h}_2 \mid \boldsymbol{h}_1, \pi_1(\boldsymbol{h}_1)\} dF_{\boldsymbol{H}_1}(\boldsymbol{h}_1), \tag{2}$$

where the right-hand side of (2) is $\text{pr}^{\pi_1,\pi_2^*}(Y \leq y)$. Let $G(\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$ denote the joint conditional distribution of $m(\boldsymbol{H}_2)$ and $c(\boldsymbol{H}_2)$ given $\boldsymbol{H}_1 = \boldsymbol{h}_1$ and $A_1 = a_1$, then $\text{pr}^{\pi_1,\pi_2^*}(Y \leq y) = E\left(I\left[y, F_\epsilon(\cdot), G\{\cdot,\cdot \mid \boldsymbol{H}_1, \pi_1(\boldsymbol{H}_1)\}\right]\right)$, where

$$I\{y, F_\epsilon(\cdot), G(\cdot,\cdot \mid \boldsymbol{h}_1, a_1)\} = \int F_\epsilon(y - u - |v|) dG(u, v \mid \boldsymbol{h}_1, a_1). \tag{3}$$

The λ-optimal regime $\boldsymbol{\pi}_\lambda^{\text{TIQ}} = \{\pi_{1,\lambda}^{\text{TIQ}}, \pi_{2,\lambda}^{\text{TIQ}}\}$ satisfies $\text{pr}^{\boldsymbol{\pi}_\lambda^{\text{TIQ}}}(Y > \lambda) \geq \text{pr}^{\boldsymbol{\pi}}(Y > \lambda)$ for all $\boldsymbol{\pi}$. That is, the distribution of Y induced by regime $\boldsymbol{\pi}_\lambda^{\text{TIQ}}$ has at least as much mass above λ as the distribution of Y induced by any other regime. It follows from the lower bound on $\text{pr}^{\pi_1,\pi_2}(Y \leq y)$ displayed in (2) that $\pi_{2,\lambda}^{\text{TIQ}}(\boldsymbol{h}_2) = \pi_2^*(\boldsymbol{h}_2) = \text{sgn}\{c(\boldsymbol{h}_2)\}$ for all $\boldsymbol{h}_2$, independent of λ and $\pi_{1,\lambda}^{\text{TIQ}}$. Henceforth, we denote $\pi_{2,\lambda}^{\text{TIQ}}$ by π_2^*. The relationship

$$\begin{aligned}\text{pr}^{\pi_1,\pi_2^*}(Y > \lambda) &= 1 - E\left(I\left[\lambda, F_\epsilon(\cdot), G\{\cdot,\cdot \mid \boldsymbol{H}_1, \pi_1(\boldsymbol{H}_1)\}\right]\right) \\ &\leq 1 - E\left[\min_{a_1} I\{\lambda, F_\epsilon(\cdot), G(\cdot,\cdot \mid \boldsymbol{H}_1, a_1)\}\right],\end{aligned} \tag{4}$$

shows that the λ-optimal first-stage rule is $\pi_{1,\lambda}^{\text{TIQ}}(\boldsymbol{h}_1) = \arg\min_{a_1} I\{\lambda, F_\epsilon(\cdot), G(\cdot,\cdot \mid \boldsymbol{h}_1, a_1)\}$. Inequality (4) holds because $I\{\lambda, F_\epsilon(\cdot), G(\cdot,\cdot \mid \boldsymbol{H}_1, a_1)\}$ is minimized over a_1 for all $\boldsymbol{H}_1$. It will be useful later on to write $\pi_{1,\lambda}^{\text{TIQ}}(\boldsymbol{h}_1) = \text{sgn}\{d(\boldsymbol{h}_1, \lambda)\}$ where

$$d(\boldsymbol{h}_1, \lambda) = I\{\lambda, F_\epsilon(\cdot), G(\cdot,\cdot \mid \boldsymbol{h}_1, -1)\} - I\{\lambda, F_\epsilon(\cdot), G(\cdot,\cdot \mid \boldsymbol{h}_1, 1)\}. \tag{5}$$

Below, we describe the general form of the TIQ-learning algorithm that can be used to estimate the λ-optimal regime. The exact algorithm depends on the choice of estimators for $m(\boldsymbol{H}_2)$, $c(\boldsymbol{H}_2)$, $F_\epsilon(\cdot)$ and $G(\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$. For example, one might posit parametric models, $m(\boldsymbol{H}_2;\boldsymbol{\beta}_{2,0})$ and $c(\boldsymbol{H}_2;\boldsymbol{\beta}_{2,1})$, for $m(\boldsymbol{H}_2)$ and $c(\boldsymbol{H}_2)$ and estimate the parameters in the model $Y = m(\boldsymbol{H}_2;\boldsymbol{\beta}_{2,0}) + c(\boldsymbol{H}_2;\boldsymbol{\beta}_{2,1}) + \epsilon$ using least squares. Alternatively, these terms could be estimated nonparametrically. We discuss possible estimators for $F_\epsilon(\cdot)$ and $G(\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$ in Sections 3.1 and 3.2. In practice, the choice of estimators should be informed by the observed data. Finally, we emphasize that if the binary threshold outcome is the terminal focus, rather than quantile optimization, a different approach (e.g., nonparametric or direct search) may be warranted. The following algorithm provides a foundation for the quantile optimization algorithm in the next section. Define $\widehat{d}(\boldsymbol{h}_1,\lambda) = I\{\lambda, \widehat{F}_\epsilon(\cdot), \widehat{G}(\cdot,\cdot \mid \boldsymbol{h}_1, -1)\} - I\{\lambda, \widehat{F}_\epsilon(\cdot), \widehat{G}(\cdot,\cdot \mid \boldsymbol{h}_1, 1)\}$.

TIQ-learning algorithm:

TIQ.1 Estimate $m(\boldsymbol{H}_2)$ and $c(\boldsymbol{H}_2)$, and denote the resulting estimates by $\widehat{m}(\boldsymbol{H}_2)$ and $\widehat{c}(\boldsymbol{H}_2)$. Given $\boldsymbol{h}_2$, estimate π_2^* using the plug-in estimator $\widehat{\pi}_2^*(\boldsymbol{h}_2) = \text{sgn}\{\widehat{c}(\boldsymbol{h}_2)\}$.

TIQ.2 Estimate $F_\epsilon(\cdot)$, the cumulative distribution function of ϵ, using the residuals $\hat{\epsilon}^Y = Y - \widehat{m}(\boldsymbol{H}_2) - A_2\widehat{c}(\boldsymbol{H}_2)$ from TIQ.1. Let $\widehat{F}_\epsilon(\cdot)$ denote this estimator.

TIQ.3 Estimate $G(\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$, the joint conditional distribution of $m(\boldsymbol{H}_2)$ and $c(\boldsymbol{H}_2)$ given $\boldsymbol{H}_1 = \boldsymbol{h}_1$ and $A_1 = a_1$. Let $\widehat{G}(\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$ denote this estimator.

TIQ.4 Given $\boldsymbol{h}_1$, estimate $\pi_{1,\lambda}^{\text{TIQ}}$ using the plug-in estimator $\widehat{\pi}_{1,\lambda}^{\text{TIQ}}(\boldsymbol{h}_1) = \text{sgn}\{\widehat{d}(\boldsymbol{h}_1,\lambda)\}$.

The TIQ-learning algorithm involves modeling $m(\boldsymbol{H}_2)$, $c(\boldsymbol{H}_2)$, the distribution function $F_\epsilon(\cdot)$, and the bivariate conditional density $G(\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$. This is more modeling than some mean-targeting algorithms. For example, Q-learning requires modeling $m(\boldsymbol{H}_2)$, $c(\boldsymbol{H}_2)$, and the conditional mean of $m(\boldsymbol{H}_2) + |c(\boldsymbol{H}_2)|$ given $\boldsymbol{H}_1, \boldsymbol{A}_1$, while IQ-learning requires modeling $m(\boldsymbol{H}_2)$, $c(\boldsymbol{H}_2)$,

the conditional mean of $m(\boldsymbol{H}_2)$, and the conditional density of $c(\boldsymbol{H}_2)$ (Laber et al., 2014a). We discuss models for the components of the TIQ-learning algorithm in Sections 3.1 and 3.2.

Remark 2.2. Standard trade-offs between parametric and nonparametric estimation apply to all terms in the TIQ-learning algorithm. In practice, the choice of estimators will likely depend on sample size and the scientific goals of the study. If the goal is to estimate a regime for immediate decision support in the clinic, then the marginal mean outcome of the estimated regime is of highest priority. Given sufficient data, it may be desirable to use nonparametric estimators in this context. However, if the goal is to inform future research and generate hypotheses for further investigation, then factors like parsimony, interpretability, and the ability to identify and test for factors associated with heterogeneous treatment response may be most important. In this context, parametric models may be preferred for components of the TIQ-learning algorithm, namely, the treatment interaction term $c(\boldsymbol{H}_2)$, while more flexible models may be specified for the nuisance function $m(\boldsymbol{H}_2)$. A parsimonious, parametric model for $c(\boldsymbol{H}_2)$ trades a restriction on the class of possible regimes for potential gains in interpretability. Note that approaches such as OWL (Zhao et al., 2012) and the methods described in Luedtke and van der Laan (2015) do not require models for $m(\boldsymbol{H}_2)$, while double robust approaches model this term only to increase efficiency and remain consistent for the optimal regime even when the model is misspecified.

Remark 2.3. Let π_1^{M} denote the first-stage decision rule of an optimal regime for the mean of Y. Then, assuming the set-up of Section 2.1, it can be shown that

$$\pi_1^{\mathrm{M}}(\boldsymbol{h}_1) = \arg\min_{a_1} \int (-u - |v|) dG(u, v \mid \boldsymbol{h}_1, a_1) = \arg\min_{a_1} \int (\lambda - u - |v|) dG(u, v \mid \boldsymbol{h}_1, a_1),$$

whereas $\pi_{1,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_1) = \arg\min_{a_1} \int F_\epsilon(\lambda - u - |v|) dG(u, v \mid \boldsymbol{h}_1, a_1)$. If $F_\epsilon(\cdot)$ is approximately linear where the conditional distribution of $\lambda - m(\boldsymbol{H}_2) - |c(\boldsymbol{H}_2)|$ given $\boldsymbol{H}_1 = \boldsymbol{h}_1$ and $A_1 = a_1$ is concentrated, $\pi_1^{\mathrm{M}}(\boldsymbol{h}_1)$ and $\pi_{1,\lambda}^{\mathrm{TIQ}}(\boldsymbol{h}_1)$ will likely agree. Thus, the difference between the mean optimal and TIQ-learning optimal regimes can be compared empirically by computing $\arg\min_{a_1} \int (-u - |v|) d\widehat{G}(u, v \mid$

$\boldsymbol{h}_{1i}, a_1)$, $\arg\min_{a_1} \int \widehat{F}_\epsilon(\lambda - u - |v|) d\widehat{G}(u, v \mid \boldsymbol{h}_{1i}, a_1)$, for each first-stage patient history $\boldsymbol{h}_{1i}$, $i = 1, \ldots, n$, and examining where these rules differ.

Remark 2.4. One approach to estimating an optimal decision rule for threshold-exceedance probabilities is discrete Q-learning (Chakraborty and Moodie, 2013). Suppose $Y = m^*(\boldsymbol{H}_2) + A_2 c^*(\boldsymbol{H}_2) + \epsilon$, where ϵ has cumulative distribution function $F_\epsilon(\cdot)$. For the binary outcome $\mathbb{1}_{Y>\lambda}$, define the second-stage Q-function, $Q_2(\boldsymbol{H}_2, A_2) = \text{pr}(\mathbb{1}_{Y>\lambda} = 1 \mid \boldsymbol{H}_2, A_2)$, and the first stage Q-function, $Q_1(\boldsymbol{H}_1, A_1) = \mathbb{E}\left[\max_{a_2} Q_2(\boldsymbol{H}_2, a_2) \mid \boldsymbol{H}_1, A_1\right]$. If these functions were known, the optimal treatment assignments for the set of observed histories $(\boldsymbol{H}_1 = \boldsymbol{h}_1, \boldsymbol{H}_2 = \boldsymbol{h}_2)$ would be $\{\arg\max_{a_1} Q_1(\boldsymbol{h}_1, a_1), \arg\max_{a_2} Q_2(\boldsymbol{h}_2, a_2)\}$. In practice, the Q-functions are unknown and the Q-learning algorithm proceeds by specifying models for them. Denote estimates of the Q-functions obtained from such models by $\widehat{Q}_2(\boldsymbol{H}_2, A_2)$ and $\widehat{Q}_1(\boldsymbol{H}_1, A_1)$. Then, the estimated optimal treatment assignments for $(\boldsymbol{H}_1 = \boldsymbol{h}_1, \boldsymbol{H}_2 = \boldsymbol{h}_2)$ are $\{\arg\max_{a_1} \widehat{Q}_1(\boldsymbol{h}_1, a_1), \arg\max_{a_2} \widehat{Q}_2(\boldsymbol{h}_2, a_2)\}$. For binary outcomes, logistic regression is often a natural model choice for Q_2. Subsequently, at the first stage one would specify a model for $\mathbb{E}\left[\max_{a_2} \widehat{Q}_2(\boldsymbol{H}_2, a_2) \mid \boldsymbol{H}_1, A_1\right]$, where the pseudo outcome $\max_{a_2} \widehat{Q}_2(\boldsymbol{H}_2, a_2)$ is bounded in $[0, 1]$. Rather than modeling this conditional expectation with linear regression, which may result in $\widehat{Q}_1$ estimates outside the interval $[0, 1]$, an alternative is to model $\mathbb{E}\left[\max_{a_2} \text{logit}\{\widehat{Q}_2(\boldsymbol{H}_2, a_2)\} \mid \boldsymbol{H}_1, A_1\right]$ using linear regression, since $\max_{a_2} \text{logit}\{\widehat{Q}_2(\boldsymbol{H}_2, a_2)\}$ takes values on the real line (Chakraborty and Moodie, 2013; Moodie et al., 2014). However, for our special case of discrete Q-learning with outcome $\mathbb{1}_{Y>\lambda}$, the logit link function is misspecified for $\text{pr}(\mathbb{1}_{Y>\lambda} = 1 \mid \boldsymbol{H}_2, A_2)$. The correct link function for the second stage Q-function is $L(u) = \lambda - F_\epsilon^{-1}(1 - u)$. When the logit link is used rather than $L(u)$, discrete Q-learning as described above and Q-learning for the continuous outcome Y perform similarly across all values of λ for the generative model for Y given above. We demonstrate this using simulation experiments in Section 4.

3. QUANTILE INTERACTIVE Q-LEARNING

Under some generative models, assigning treatment according to a mean-optimal regime leads to higher average outcomes at the expense of higher variability, negatively affecting patients with outcomes in the lower quantiles of the induced distribution of Y. We demonstrate this using simulated examples in Section 4. Define the τ^{th} quantile of the distribution of Y induced by regime $\boldsymbol{\pi}$ as $q^{\boldsymbol{\pi}}(\tau) = \inf\{y : \text{pr}^{\pi_1,\pi_2}(Y \leq y) \geq \tau\}$. The goal of Quantile Interactive Q-learning (QIQ-learning) is to estimate a pair of decision rules, $\boldsymbol{\pi}_\tau^{\text{QIQ}} = \{\pi_{1,\tau}^{\text{QIQ}}, \pi_{2,\tau}^{\text{QIQ}}\}$, that maximize $q^{\boldsymbol{\pi}}(\tau)$ over $\boldsymbol{\pi}$ for a fixed, prespecified τ. QIQ-learning is similar to TIQ-learning, but the optimal first-stage rule is complicated by the inversion of the distribution function to obtain quantiles of Y under a given regime. When the variance of Y is independent of A_2, the QIQ-learning second-stage optimal decision is $\pi_{2,\tau}^{\text{QIQ}}(\boldsymbol{h}_2) = \pi_2^*(\boldsymbol{h}_2) = \text{sgn}\{c(\boldsymbol{h}_2)\}$, independent of τ and $\pi_{1,\tau}^{\text{QIQ}}$; details are provided in Section 4 of the supplemental material. Denote $\pi_{2,\tau}^{\text{QIQ}}$ by π_2^*.

Next we characterize $\pi_{1,\tau}^{\text{QIQ}}$, which will motivate an algorithm for calculating it. Let $d(\boldsymbol{h}_1, y)$ be as in (5), and define $\Gamma(\boldsymbol{h}_1, y) \triangleq \text{sgn}\{d(\boldsymbol{h}_1, y)\}$. Then $\Gamma(\boldsymbol{h}_1, y) = \pi_{1,\lambda}^{\text{TIQ}}(\boldsymbol{h}_1)|_{\lambda=y}$ is the optimal first-stage decision rule of TIQ-learning at $\lambda = y$. We have introduced the new notation to emphasize the dependence on y. Next, define the optimal τ^{th} quantile

$$y_\tau^* \triangleq \inf\left\{y : \text{pr}^{\Gamma(\cdot,y),\pi_2^*}(Y \leq y) \geq \tau\right\}, \tag{6}$$

which we study further in the remainder of this section.

Lemma 7.7 of the supplemental material proves that $\lim_{y\to\infty(-\infty)} \text{pr}^{\Gamma(\cdot,y),\pi_2^*}(Y \leq y) = 1(0)$, so that y_τ^* is defined for all $\tau \in (0,1)$. For each $y \in \mathbb{R}$,

$$\begin{aligned}
\text{pr}^{\pi_1,\pi_2^*}(Y \leq y) &= E\left(I\left[y, F_\epsilon(\cdot), G\{\cdot,\cdot \mid \boldsymbol{H}_1, \pi_1(\boldsymbol{H}_1)\}\right]\right) \\
&\geq E\left(I\left[y, F_\epsilon(\cdot), G\{\cdot,\cdot \mid \boldsymbol{H}_1, \Gamma(\boldsymbol{H}_1, y)\}\right]\right) \\
&= \text{pr}^{\Gamma(\cdot,y),\pi_2^*}(Y \leq y),
\end{aligned}$$

where $I(\cdot,\cdot,\cdot)$ is defined in (3). The last equality follows because $\Gamma(\boldsymbol{H}_1, y)$ minimizes $E\left(I\left[y, F_\epsilon(\cdot),\right.\right.$

$G\{\cdot,\cdot \mid \boldsymbol{H}_1, a_1\}$

with respect to a_1. Hence, $\left\{y : \text{pr}^{\Gamma(\cdot,y),\pi_2^*}(Y \le y) \ge \tau\right\} \subseteq \left\{y : \text{pr}^{\pi_1,\pi_2^*}(Y \le y) \ge \tau\right\}$, and taking the infimum on both sides gives the upper bound

$$y_\tau^* \ge q^{\pi_1,\pi_2^*}(\tau) \text{ for all } \pi_1. \tag{7}$$

Thus, a first-stage decision rule π_1 is optimal if it induces a τ^{th} quantile equal to the upper bound y_τ^* when treatments are subsequently assigned according to π_2^*, i.e., if $q^{\pi_1,\pi_2^*}(\tau) = y_\tau^*$.

We now discuss conditions that guarantee existence of a π_1 such that $q^{\pi_1,\pi_2^*}(\tau) = y_\tau^*$ and derive its form. The quantile obtained under regime $\pi = \{\Gamma(\cdot,y),\pi_2^*\}$ is

$$f(y) \triangleq q^{\Gamma(\cdot,y),\pi_2^*}(\tau) = \inf\left\{\widetilde{y} : \text{pr}^{\Gamma(\cdot,y),\pi_2^*}(Y \le \widetilde{y}) \ge \tau\right\}. \tag{8}$$

Thus, because it is a quantile and the bound in (7) applies, $\text{pr}^{\Gamma(\cdot,y),\pi_2^*}\{Y \le f(y)\} \ge \tau$, and $f(y) = q^{\Gamma(\cdot,y),\pi_2^*}(\tau) \le y_\tau^*$ for all y. Our main results depend on the following lemma, which is proved in the supplemental material.

Lemma 3.1.

(*A*) $y < y_\tau^*$ *implies* $y < f(y) \le y_\tau^*$; (9)

(*B*) $f(y_\tau^{*-}) \triangleq \lim_{\delta \downarrow 0} f(y_\tau^* - \delta) = y_\tau^*$; (10)

(*C*) $f(y_\tau^*) \le y_\tau^*$ *with strict inequality if there exists* $\delta > 0$ *such that*

$pr^{\Gamma(\cdot,y_\tau^*),\pi_2^*}(Y \le y_\tau^* - \delta) \ge \tau$; (11)

(*D*) *If* $F_\epsilon(\cdot)$ *is continuous and strictly increasing, then* $f(y_\tau^*) = y_\tau^*$. (12)

It follows from (*B*) that $f(y_\tau^*) = y_\tau^*$ if and only if $f(y)$ is left continuous at $y = y_\tau^*$, and part (*D*) is a sufficient condition guaranteeing left-continuity of $f(y)$ at y_τ^*. In this case, the optimal first-stage rule is $\pi_{1,\tau}^{\text{QIQ}}(\boldsymbol{h}_1) = \Gamma(\boldsymbol{h}_1, y_\tau^*)$, i.e., $q^{\Gamma(\cdot,y_\tau^*),\pi_2^*}(\tau) = y_\tau^*$. The condition stated in (*D*) is commonly satisfied, e.g., when the density of ϵ has positive support on the entire real line. If $f(y)$ is not

left continuous at y_τ^*, and thus $f(y_\tau^*) < y_\tau^*$, in light of (10) we can always approach the optimal policy via a sequence of regimes of the form $\{\Gamma(\cdot, y_\tau^* - \delta_n), \pi_2^*\}$, where δ_n decreases to 0. If the underlying distributions of the histories and Y were known, the following algorithm produces an optimal regime.

Population-level algorithm to find $\pi_{1,\tau}^{\text{QIQ}}$:

1. Compute y_τ^* from (6) and $f(y_\tau^*)$ from (8).

2. a. If $f(y_\tau^*) = y_\tau^*$, $\pi_{1,\tau}^{\text{QIQ}}(\boldsymbol{h}_1) = \Gamma(\boldsymbol{h}_1, y_\tau^*)$ is optimal as it attains the quantile y_τ^*.

   b. If $f(y_\tau^*) < y_\tau^*$, $\pi_{1,\tau}^{\text{QIQ}}(\boldsymbol{h}_1) = \lim_{\delta \downarrow 0} \Gamma(\boldsymbol{h}_1, y_\tau^* - \delta)$ is optimal. Note that this rule can be written in closed form as $\lim_{\delta \downarrow 0} \Gamma(\boldsymbol{h}_1, y_\tau^* - \delta) = \text{sgn}\{d^-(\boldsymbol{h}_1, y_\tau^*)\}$, where we define $d^-(\boldsymbol{h}_1, y_\tau^*) = I_\epsilon(a_1)|_{a_1=-1} - I_\epsilon(a_1)|_{a_1=1}$ and $I_\epsilon(a_1) = \int \text{pr}(\epsilon < y_\tau^* - u - |v|) dG(u, v \mid \boldsymbol{h}_1, a_1)$.

In practice, the generative model is not known, but the population-level algorithm suggests an estimator of $\pi_{1,\tau}^{\text{QIQ}}$. The following QIQ-learning algorithm can be used to estimate an optimal first-stage decision rule. The exact algorithm depends on the choice of estimators for $F_\epsilon(\cdot)$ and $G(\cdot, \cdot \mid \boldsymbol{h}_1, a_1)$; several options are presented in Sections 3.1 and 3.2, but the choice should be data-driven; see, e.g., Remark 2.2.

QIQ-learning algorithm:

QIQ.1 Follow TIQ.1 – TIQ.3 of the TIQ-learning algorithm in Section 2.1.

QIQ.2 With $I(\cdot, \cdot, \cdot)$ as in (3) and first-stage patient histories $\boldsymbol{h}_{1i}$, estimate y_τ^* using

$$\widehat{y}_\tau^* \triangleq \inf\left(y : \frac{1}{n} \sum_{i=1}^{n} I\left[y, \widehat{F}_\epsilon(\cdot), \widehat{G}\{\cdot, \cdot \mid \boldsymbol{h}_{1i}, \widehat{\Gamma}(\boldsymbol{h}_{1i}, y)\} \right] \geq \tau \right).$$

QIQ.3 Estimate $f(y_\tau^*)$ using

$$\widehat{f(y_\tau^*)} \triangleq \inf\left(y : \frac{1}{n} \sum_{i=1}^{n} I\left[y, \widehat{F}_\epsilon(\cdot), \widehat{G}\{\cdot, \cdot \mid \boldsymbol{h}_{1i}, \widehat{\Gamma}(\boldsymbol{h}_{1i}, \widehat{y}_\tau^*)\} \right] \geq \tau \right).$$

QIQ.4 a. If $\widehat{f}(\widehat{y}^*_\tau) = \widehat{y}^*_\tau$, then $\widehat{\pi}^{\text{QIQ}}_{1,\tau}(\boldsymbol{h}_1) = \widehat{\Gamma}(\boldsymbol{h}_1, \widehat{y}^*_\tau)$ is an estimated optimal first-stage decision rule because it attains the estimated optimal quantile, $\widehat{y}^*_\tau$, when treatments are subsequently assigned according to π^*_2 at the second stage.

b. If $\widehat{f}(\widehat{y}^*_\tau) < \widehat{y}^*_\tau$, then the first-stage rule $\widehat{\pi}^{\text{QIQ}}_{1,\tau}(\boldsymbol{h}_1) = \widehat{\Gamma}(\boldsymbol{h}_1, \widehat{y}^*_\tau - \delta)$, $\delta > 0$, results in the estimated quantile $\widehat{f}(\widehat{y}^*_\tau - \delta)$, which satisfies $\widehat{y}^*_\tau - \delta < \widehat{f}(\widehat{y}^*_\tau - \delta) \leq \widehat{y}^*_\tau$. By choosing δ arbitrarily small, this estimated quantile will be arbitrarily close to the estimated optimal quantile $\widehat{y}^*_\tau$.

To complete the TIQ- and QIQ-learning algorithms, we provide specific estimators $F_\epsilon(\cdot)$ and $G(\cdot, \cdot \mid \boldsymbol{h}_1, a_1)$ in the next two sections. We suggest estimators that are likely to be useful in practice, but our list is not exhaustive. An advantage of TIQ- and QIQ-learning is that they involve modeling only smooth transformations of the data; these are standard, well-studied modeling problems in the statistics literature.

3.1 Working models for $F_\epsilon(\cdot)$

Both TIQ- and QIQ-learning require estimation of the distribution function of the second-stage error, ϵ. We suggest two estimators that are useful in practice. The choice between them can be guided by inspection of the residuals from the second-stage regression.

Normal Scale Model.

The normal scale estimator for $F_\epsilon(\cdot)$ is $\widehat{F}^N_\epsilon(z) \triangleq \Phi\left(z/\widehat{\sigma}_\epsilon\right)$, where $\Phi(\cdot)$ denotes the standard normal distribution function and $\widehat{\sigma}_\epsilon$ is the standard deviation of the second-stage residuals, $\widehat{e}^Y_i \triangleq Y_i - m(\boldsymbol{H}_{2i}) - A_{2i}c(\boldsymbol{H}_{2i})$, $i = 1, \ldots, n$. If it is thought that σ_ϵ depends on $(\boldsymbol{H}_2, A_2)$, flexibility can be gained by assuming a heteroskedastic variance model (Carroll and Ruppert, 1988), i.e., by assuming $F_\epsilon(z) = \Phi\left\{z/\sigma_\epsilon(\boldsymbol{H}_2, A_2)\right\}$ for some unknown function $\sigma_\epsilon(\boldsymbol{h}_2, a_2)$. Given an estimator $\widehat{\sigma}_\epsilon(\boldsymbol{h}_2, a_2)$ of $\sigma_\epsilon(\boldsymbol{h}_2, a_2)$, an estimator of $F_\epsilon(\cdot)$ is $\widehat{F}^N_\epsilon(z) \triangleq \Phi\left\{z/\widehat{\sigma}_\epsilon(\boldsymbol{H}_2, A_2)\right\}$. We discuss variance modeling techniques in the next section.

Nonparametric Model. For more flexibility, a non- or semi-parametric estimator for $F_\epsilon(\cdot)$ can be used. In the homogeneous variance case, a nonparametric estimator of $F_\epsilon(\cdot)$ is the empirical distribution of the residuals, $\widehat{F}^E_\epsilon(z) \triangleq n^{-1}\sum_{i=1}^n \mathbb{1}(\widehat{e}^Y_i \leq z)$. In the heterogeneous variance case, one can assume a non- or semi-parametric scale model $F_{\epsilon|\boldsymbol{H}_2,A_2}(z|\boldsymbol{H}_2 = \boldsymbol{h}_2, A_2 = a_2) = F_0\{z/\sigma_\epsilon(\boldsymbol{h}_2, a_2)\}$, where $F_0(\cdot)$ is an unspecified distribution function. Given an estimator $\widehat{\sigma}_\epsilon(\boldsymbol{h}_2, a_2)$ of $\sigma_\epsilon(\boldsymbol{h}_2, a_2)$, an estimator of $F_{\epsilon|\boldsymbol{H}_2,A_2}(z|\boldsymbol{H}_2 = \boldsymbol{h}_2, A_2 = a_2)$ is $\widehat{F}^E_\epsilon(z|\boldsymbol{H}_2 = \boldsymbol{h}_2, A_2 = a_2) = n^{-1}\sum_{i=1}^n \mathbb{1}\left\{\widetilde{e}^Y_i \leq z/\widehat{\sigma}_\epsilon(\boldsymbol{h}_2, a_2)\right\}$, where $\widetilde{e}^Y_i = \widehat{e}^Y_i/\widehat{\sigma}_\epsilon(\boldsymbol{H}_{2i}, A_{2i})$. Standard residual diagnostic techniques, e.g., a normal quantile-quantile plot, can be used to determine whether a normal assumption seems plausible for the observed data.

3.2 Working models for $G(\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$

In addition to a model for $F_\epsilon(\cdot)$, TIQ- and QIQ-learning require models for the bivariate conditional density of $m(\boldsymbol{H}_2)$ and $c(\boldsymbol{H}_2)$ given $\boldsymbol{H}_1$ and A_1. A useful strategy is to first model the conditional mean and variance functions of $m(\boldsymbol{H}_2)$ and $c(\boldsymbol{H}_2)$ and then estimate the joint distribution of their standardized residuals. Define these standardized residuals as

$$e^m = \frac{m(\boldsymbol{H}_2) - \mu_m(\boldsymbol{H}_1, A_1)}{\sigma_m(\boldsymbol{H}_1, A_1)}, \qquad e^c = \frac{c(\boldsymbol{H}_2) - \mu_c(\boldsymbol{H}_1, A_1)}{\sigma_c(\boldsymbol{H}_1, A_1)},$$

where $\mu_m(\boldsymbol{H}_1, A_1) \triangleq E\{m(\boldsymbol{H}_2) \mid \boldsymbol{H}_1, A_1\}$ and $\sigma^2_m(\boldsymbol{H}_1, A_1) \triangleq E[\{m(\boldsymbol{H}_2) - \mu_m(\boldsymbol{H}_1, A_1)\}^2 \mid \boldsymbol{H}_1, A_1]$. The mean and variance functions of $c(\boldsymbol{H}_2)$ are defined similarly: $\mu_c(\boldsymbol{H}_1, A_1) \triangleq E\{c(\boldsymbol{H}_2)|\boldsymbol{H}_1, A_1\}$, and $\sigma^2_c(\boldsymbol{H}_1, A_1) \triangleq E[\{c(\boldsymbol{H}_2) - \mu_c(\boldsymbol{H}_1, A_1)\}^2|\boldsymbol{H}_1, A_1]$. In simulations, we use parametric mean and variance models for μ_m, σ^2_m, μ_c, and σ^2_c, and we estimate the joint distribution of e^m and e^c using a Gaussian copula. Alternatively, the joint residual distribution could be modelled parametrically, e.g., with a multivariate normal model; or nonparametrically, e.g., using a bivariate kernel density estimator (Silverman, 1986, Ch. 4). The Gaussian copula is used in the simulations in Section 4, and results are provided using a bivariate kernel estimator in the supplemental material. Common exploratory analysis techniques can be used to interactively guide the choice of estimator for

$G(\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$. In simulated experiments described in the supplemental material, a bivariate kernel density estimator was competitive with a correctly specified Gaussian copula model with sample sizes as small as $n = 100$. Using parametric mean and variance modeling, the following steps would be substituted in Step TIQ.3 of the TIQ-learning algorithm.

Mean and Variance Modeling.

3.1 Compute $\widehat{\boldsymbol{\theta}}_m \triangleq \arg\min_{\boldsymbol{\theta}_m} \sum_{i=1}^n \{\widehat{m}(\boldsymbol{H}_{2i}) - \mu_m(\boldsymbol{H}_{1i}, A_{1i}; \boldsymbol{\theta}_m)\}^2$ and the resulting estimator $\mu_m(\boldsymbol{H}_1, A_1; \widehat{\boldsymbol{\theta}}_m)$ of the mean function $\mu_m(\boldsymbol{H}_1, A_1)$.

3.2 Use the estimated mean function from Step 3.1 to obtain

$$\widehat{\boldsymbol{\gamma}}_m \triangleq \arg\min_{\gamma_m} \sum_{i=1}^n \left[\left\{ \widehat{m}(\boldsymbol{H}_{2i}) - \mu_m(\boldsymbol{H}_{1i}, A_{1i}; \widehat{\boldsymbol{\theta}}_m) \right\}^2 - \sigma_m^2(\boldsymbol{H}_{1i}, A_{1i}; \boldsymbol{\gamma}_m) \right]^2,$$

and subsequently the estimator $\sigma_m^2(\boldsymbol{H}_1, A_1; \widehat{\boldsymbol{\gamma}}_m)$ of $\sigma_m^2(\boldsymbol{H}_1, A_1)$. One choice for $\sigma_m(\boldsymbol{h}_1, a_2; \gamma_m)$ is a log-linear model, which may include non-linear basis terms.

3.3 Repeat Steps 3.1 and 3.2 to obtain estimators $\mu_c(\boldsymbol{H}_1, A_1; \widehat{\boldsymbol{\theta}}_c)$ and $\sigma_c(\boldsymbol{H}_1, A_1; \widehat{\boldsymbol{\gamma}}_c)$.

3.4 Compute standardized residuals $\widehat{e}_i^m$ and $\widehat{e}_i^c$, $i = 1, ..., n$, as

$$\widehat{e}_i^m = \frac{\widehat{m}(\boldsymbol{H}_{2i}) - \mu_m(\boldsymbol{H}_{1i}, A_{1i}; \widehat{\boldsymbol{\theta}}_m)}{\sigma_m(\boldsymbol{H}_{1i}, A_{1i}; \widehat{\boldsymbol{\gamma}}_m)}, \qquad \widehat{e}_i^c = \frac{\widehat{c}(\boldsymbol{H}_{2i}) - \mu_c(\boldsymbol{H}_{1i}, A_{1i}; \widehat{\boldsymbol{\theta}}_c)}{\sigma_c(\boldsymbol{H}_{1i}, A_{1i}; \widehat{\boldsymbol{\gamma}}_c)}.$$

Then, $\widehat{e}_i^m$ and $\widehat{e}_i^c$, $i = 1, ..., n$, can be used to estimate the joint distribution of the standardized residuals. Samples drawn from this distribution can be transformed back to samples from $\widehat{G}(\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$ to estimate the integral $I\left\{y, \widehat{F}_\epsilon(\cdot), \widehat{G}(\cdot,\cdot \mid \boldsymbol{h}_1, a_1)\right\}$ with a Monte Carlo average.

3.3 Theoretical results

The following assumptions are used to establish consistency of the threshold exceedance probability and quantile that result from applying the estimated TIQ- and QIQ-learning optimal regimes,

respectively. For each $\boldsymbol{h}_1, a_1$, and $\boldsymbol{h}_2$:

A1. the method used to estimate $m(\cdot)$ and $c(\cdot)$ results in estimators $\widehat{m}(\boldsymbol{h}_2)$ and $\widehat{c}(\boldsymbol{h}_2)$ that converge in probability to $m(\boldsymbol{h}_2)$ and $c(\boldsymbol{h}_2)$, respectively;

A2. $F_\epsilon(\cdot)$ is continuous, $\widehat{F}_\epsilon(\cdot)$ is a cumulative distribution function, and $\widehat{F}_\epsilon(y)$ converges in probability to $F_\epsilon(y)$ uniformly in y;

A3. $\int |d\widehat{G}(u, v \mid \boldsymbol{h}_1, a_1) - dG(u, v \mid \boldsymbol{h}_1, a_1)|$ converges to zero in probability;

A4. $n^{-1} \sum_{i=1}^{n} \int |d\widehat{G}(u, v \mid \boldsymbol{H}_{1i}, a_1) - dG(u, v \mid \boldsymbol{H}_{1i}, a_1)|$ converges to zero in probability.

In the simulation experiments in Section 4 and data example in Section 5, we use linear working models for $m(\cdot)$ and $c(\cdot)$ that are estimated using least squares. Thus, A1 is satisfied under usual regularity conditions. When ϵ is continuous, assumption A2 can be satisfied by specifying $\widehat{F}_\epsilon(\cdot)$ as the empirical distribution function. If for each fixed $\boldsymbol{h}_1$ and a_1, $dG(\cdot, \cdot \mid \boldsymbol{h}_1, a_1)$ is a density and $d\widehat{G}(\cdot, \cdot \mid \boldsymbol{h}_1, a_1)$ a pointwise consistent estimator, then A3 is satisfied (Glick, 1974). Theorems 3.2 and 3.3 are proved in the supplemental material.

Theorem 3.2. *(Consistency of TIQ-learning) Assume A1–A3 and fix $\lambda \in \mathbb{R}$. Then, $pr^{\widehat{\boldsymbol{\pi}}_\lambda^{TIQ}}(Y > \lambda)$ converges in probability to $pr^{\boldsymbol{\pi}_\lambda^{TIQ}}(Y > \lambda)$, where $\widehat{\boldsymbol{\pi}}_\lambda^{TIQ} = (\widehat{\pi}_{1,\lambda}^{TIQ}, \widehat{\pi}_2^*)$.*

Theorem 3.3. *(Consistency of QIQ-learning) Assume A1–A4. Then, $q^{\widehat{\boldsymbol{\pi}}_\tau^{QIQ}}(\tau)$ converges in probability to y_τ^* for any fixed τ, where $\widehat{\boldsymbol{\pi}}_\tau^{QIQ} = (\widehat{\Gamma}(\cdot, \widehat{y}_\tau^*), \widehat{\pi}_2^*)$.*

4. SIMULATION EXPERIMENTS

We compare the performance of our estimators to binary Q-learning (Chakraborty and Moodie, 2013), Q-learning, and mean-optimal IQ-learning (Laber et al., 2014a) for a range of data generative models. Gains are achieved in terms of the proportion of the distribution of Y that exceeds the constant threshold λ and the τ^{th} quantile for several values of λ and τ. The data are generated

using the model

$$
\begin{aligned}
&\boldsymbol{X}_1 \sim \text{Norm}(\mathbf{1}_2, \boldsymbol{\Sigma}), && A_1, A_2 \sim \text{Unif}\{-1, 1\}^2, && \boldsymbol{H}_1 = (1, \boldsymbol{X}_1^\intercal)^\intercal, \\
&\eta_{\boldsymbol{H}_1, A_1} = \exp\{\tfrac{C}{2}(\boldsymbol{H}_1^\intercal \boldsymbol{\gamma}_0 + A_1 \boldsymbol{H}_1^\intercal \boldsymbol{\gamma}_1)\}, && \boldsymbol{\xi} \sim \text{Norm}(\mathbf{0}_2, \mathbf{I}_2), && \boldsymbol{X}_2 = \boldsymbol{B}_{A_1} \boldsymbol{X}_1 + \eta_{\boldsymbol{H}_1, A_1} \boldsymbol{\xi}, \\
&\boldsymbol{H}_2 = (1, \boldsymbol{X}_2^\intercal)^\intercal, && \epsilon \sim \text{Norm}(0, 1), && Y = \boldsymbol{H}_2^\intercal \boldsymbol{\beta}_{2,0} + A_2 \boldsymbol{H}_2^\intercal \boldsymbol{\beta}_{2,1} + \epsilon,
\end{aligned}
$$

where $\mathbf{1}_p$ is a $p \times 1$ vector of 1s, $\mathbf{I}_q$ is the $q \times q$ identify matrix, and $C \in [0, 1]$ is a constant. The matrix $\boldsymbol{\Sigma}$ is a correlation matrix with off-diagonal $\rho = 0.5$. The 2×2 matrix $\boldsymbol{B}_{A_1}$ equals

$$
\boldsymbol{B}_{A_1=1} = \begin{pmatrix} -0.1 & -0.1 \\ 0.1 & 0.1 \end{pmatrix}, \qquad \boldsymbol{B}_{A_1=-1} = \begin{pmatrix} 0.5, & -0.1 \\ -0.1 & 0.5 \end{pmatrix}.
$$

The remaining parameters are $\boldsymbol{\gamma}_0 = (1, 0.5, 0)^\intercal$, $\boldsymbol{\gamma}_1 = (-1, -0.5, 0)^\intercal$, $\boldsymbol{\beta}_{2,0} = (0.25, -1, 0.5)^\intercal$, and $\boldsymbol{\beta}_{2,1} = (1, -0.5, -0.25)^\intercal$, which were chosen to ensure that the mean-optimal treatment produced a more variable response for some patients.

4.1 TIQ-learning Simulation Results

Results are based on $J = 1,000$ generated data sets. For each, we estimate the TIQ-, IQ-, binary Q-learning, and Q-learning policies using a training set of size $n = 250$ and compare the results using a test set of size $N = 10,000$. The normal scale model is used to estimate $F_\epsilon(\cdot)$, which is correctly specified for the generative model above. The Gaussian copula model discussed in Section 2.4 is also correctly specified and is used as the estimator for $G(\cdot, \cdot \mid \boldsymbol{h}_1, a_1)$. Results using a bivariate kernel estimator for $G(\cdot, \cdot \mid \boldsymbol{h}_1, a_1)$ are presented in the supplemental material.

To study the performance of the TIQ-learning algorithm, we compare values of the cumulative distribution function of the final response when treatment is assigned according to the estimated TIQ-learning, IQ-learning, binary Q-learning, and Q-learning regimes. Define $\text{pr}^{\widehat{\boldsymbol{\pi}}_j}(Y > \lambda)$ to be the true probability that Y exceeds λ given treatments are assigned according to $\widehat{\boldsymbol{\pi}}_j = (\widehat{\pi}_{1j}, \widehat{\pi}_{2j})$, the regime estimated from the j^{th} generated data set. For threshold values $\lambda = -2, 2, 4$, we estimate

$\text{pr}^{\boldsymbol{\pi}}(Y > \lambda)$ using $\sum_{j=1}^{J} \widehat{\text{pr}}^{\widehat{\boldsymbol{\pi}}_j}(Y > \lambda)/J$, where $\widehat{\text{pr}}^{\widehat{\boldsymbol{\pi}}_j}(Y > \lambda)$ is an estimate of $\text{pr}^{\widehat{\boldsymbol{\pi}}_j}(Y > \lambda)$ obtained by calculating the proportion of test patients consistent with regime $\widehat{\boldsymbol{\pi}}_j$ whose observed Y values are greater than λ. Thus, our estimate is an average over training data sets and test set observations. In terms of the proportion of distribution mass above λ, results for $\lambda = -2$ and 4 in Figure 1 show a clear advantage of TIQ-learning for higher values of C, the degree of heteroskedasticity in the second-stage covariates $\boldsymbol{X}_2$. As anticipated by Remark 2.3 in Section 2.1, all methods perform similarly when $\lambda = 2$.

Figure 2 illustrates how the optimal first-stage treatment for a test set of 1,000 individuals changes as λ varies. Results are shown for $C = 0.5$. The true optimal treatments displayed in the left plot show a distinct shift from treating most of the population with $A_1 = 1$ to $A_1 = -1$ as λ increases from -4 to 4. The TIQ-learning estimated optimal treatments displayed in the middle plot are averaged over 100 Monte Carlo iterations and closely resemble the true policies on the left. Although the estimated Q-learning regime does not depend on λ, it is plotted for each λ value to aid visual comparison. The first-stage treatments recommended by Q-learning differ the most from the true optimal treatments when $\lambda = 4$, corroborating the results for $C = 0.5$ in Figure 1. The rightmost panel of Figure 2 are the results from binary Q-learning with the binary outcome defined as $\mathbb{1}_{Y>\lambda}$. While there appears to be a slight deviation in the results from mean-optimal Q-learning for λ values 3 and 4, overall the resulting policies are similar to mean-optimal Q-learning and do not recover the true optimal treatments on average.

4.2 QIQ-learning Simulations

To study the performance of the QIQ-learning algorithm, we compare quantiles of Y when the population is treated according to the regimes estimated by QIQ-learning, IQ-learning, and Q-learning. A smaller test set of size $N = 5,000$ was used in this section to reduce computation time. Define $q^{\widehat{\boldsymbol{\pi}}_j}(\tau)$ to be the true τ^{th} quantile of the distribution of Y given treatments are assigned according to $\widehat{\boldsymbol{\pi}}_j = (\widehat{\pi}_{1j}, \widehat{\pi}_{2j})$, the regime estimated from the j^{th} generated data set. For $\tau = 0.1, 0.5, 0.75$, we

estimate $q^{\pi}(\tau)$ using $\sum_{j=1}^{J} \widehat{q^{\pi}}_j(\tau)/J$, where $\widehat{q^{\pi}}_j(\tau)$ is an estimate of $q^{\widehat{\pi}_j}(\tau)$ obtained by calculating the τ^{th} quantile of the subgroup of test patients consistent with regime $\widehat{\pi}_j$. The generative model and all other parameter settings used here are the same as those in the previous section. For our generative model, the condition of Lemma 3.1 is satisfied, so the true optimal regime is attained asymptotically.

The results in Figure 3 indicate that the lowest quantile, $\tau = 0.1$, suffers under the Q-learning regime as heterogeneity in the second-stage histories increases, measured by the scaling constant C. In contrast, quantiles of the QIQ-learning estimated regimes for $\tau = 0.1$ remain constant across the entire range of C. When $\tau = 0.5$, all methods perform similarly; for some C, IQ- and Q-learning outperform QIQ-learning. This is not surprising because all models used to generate the data were symmetric. Thus, maximizing the mean of Y gives similar results to maximizing the median.

Next we study QIQ-learning when the first stage errors are skewed. The generative model and parameter settings used here are the same as those used previously except that

$$\boldsymbol{X}_2 = \boldsymbol{B}_{A_1}\boldsymbol{X}_1 + \eta_{\boldsymbol{H}_1,A_1}\boldsymbol{\xi}, \qquad \eta_{\boldsymbol{H}_1,A_1} = \exp\{\tfrac{1}{2}(\boldsymbol{H}_1^{\intercal}\boldsymbol{\gamma}_0 + A_1\boldsymbol{H}_1^{\intercal}\boldsymbol{\gamma}_1)\},$$
$$(10C+1) + \sqrt{2(10C+1)}\boldsymbol{\xi} \sim \chi^2_{df=10C+1},$$

where $C \in [0,1]$ is a constant that reflects the degree of skewness in the first-stage errors, $\boldsymbol{\xi}$. Smaller values of C correspond to heavier skew.

Results are averaged over $J = 100$ generated data sets; for each, we estimate the QIQ-, IQ-, and Q-learning policies and compare the results using a test set of size $N = 10,000$. The training sample size for each iteration is $n = 500$. The normal scale model is used to estimate $F_{\epsilon}(\cdot)$, which is correctly specified. A bivariate kernel density estimator is used to estimate $G(\cdot,\cdot \mid \boldsymbol{h}_1, a_1)$. As before, we compare quantiles of the final response when treatment is assigned according to the estimated QIQ-learning, IQ-learning, and Q-learning regimes, and the results are given in Figure

4. QIQ-learning demonstrates an advantage over the mean-optimal methods for all three quantiles and almost uniformly across the degree of skewness of the first-stage errors.

5. STAR*D ANALYSIS

The Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial (Fava et al., 2003; Rush et al., 2004) is a four-stage Sequential Multiple Assignment Randomized Trial (Lavori and Dawson, 2004; Murphy, 2005a) studying personalized treatment strategies for patients with major depressive disorder. Depression is measured by the Quick Inventory of Depressive Symptomatology (QIDS) score, a one-number summary score that takes integer values 0 to 27. Lower scores indicate fewer depression symptoms. Remission is defined as QIDS $\leq$ 5. Previous attempts to estimate optimal dynamic treatment regimes from this data have used the criteria, "maximize end-of-stage-two QIDS," (see, for example, Schulte et al., 2012; Laber et al., 2014a) a surrogate for the primary aim of helping patients achieve remission. We illustrate TIQ-learning by estimating an optimal regime that maximizes the probability of remission for each patient, directly corresponding to the primary clinical goal.

The first stage, which we will henceforth refer to as baseline, was non-randomized with each patient receiving Citalopram, a drug in the class of Selective Serotonin Reuptake Inhibitors (SSRIs). We use a subset of the STAR*D data from the first two randomized stages, and refer to the original trial levels 2 and 3 as "stage one" and "stage two." Before each randomization, patients specified a preference to "switch" or "augment" their current treatment strategy and were then randomized to one of multiple options within their preferred category. In addition, patients who achieved remission in any stage exited the study. To keep our illustration of TIQ-learning concise, we restrict our analysis to the subset of patients who who preferred the "switch" strategy at both stages. We note that this subgroup is not identifiable at baseline because patient preferences depend on the assigned treatment and subsequent response at each stage. Our motivation for this restriction is to mimic a two-stage SMART where treatments are randomized at both stages, thus

simplifying our illustration. At stage one, our binary treatment variable is "SSRI," which includes only Sertraline, versus "non-SSRI," which includes both Bupropion and Venlafaxine. At stage two we compare Mirtazapine and Nortriptyline which are both non-SSRIs. In the patient subgroup considered in our analysis, treatments were randomized at both stages.

All measured QIDS scores are recoded as 27− QIDS so that higher scores correspond to fewer depression symptoms. After recoding, remission corresponds to QIDS > 21. Thus, TIQ-learning with $\lambda = 21$ maximizes the probability of remission for all patients. In general, QIDS was recorded during clinic visits at weeks 2, 4, 6, 9, and 12 in each stage, although some patients with inadequate response moved on to the next stage before completing all visits. We summarize longitudinal QIDS trajectories from the baseline stage and stage one by averaging over the total number of QIDS observations in the given stage. Variables used in our analysis are listed in Table 1.

We describe all models used in the analysis below.

At the second stage, we assume the linear working model $Y = \boldsymbol{H}_{2,0}^{\intercal}\boldsymbol{\beta}_{2,0} + A_2\boldsymbol{H}_{2,1}^{\intercal}\boldsymbol{\beta}_{2,1} + \epsilon$, where $\boldsymbol{H}_{2,0} = \boldsymbol{H}_{2,1} = (1,\ \texttt{qids1},\ \texttt{slope1},\ \texttt{A1})^{\intercal}$, $E(\epsilon) = 0$, $\text{var}(\epsilon) = \sigma^2$, and ϵ is independent of $\boldsymbol{H}_2$ and A_2. We fit this model using least squares. A normal qq-plot of the residuals from the previous regression step indicates slight deviation from normality, so we use the non-parametric estimator of $F_\epsilon(\cdot)$ given in Section 2.4. Next, we estimate the conditional mean and variance functions of $m(\boldsymbol{H}_2) \triangleq \boldsymbol{H}_{2,0}^{\intercal}\boldsymbol{\beta}_{2,0}$ and $c(\boldsymbol{H}_2) \triangleq \boldsymbol{H}_{2,1}^{\intercal}\boldsymbol{\beta}_{2,1}$ following steps described in Section 2.4. For the mean functions, we take $\boldsymbol{H}_{1,0} = \boldsymbol{H}_{1,1} = (1, \boldsymbol{X}_1^{\intercal})^{\intercal}$ with $\boldsymbol{X}_1 = (\texttt{qids0},\ \texttt{slope0})^{\intercal}$ and use working models of the form $E\{k(\boldsymbol{H}_2) \mid \boldsymbol{X}_1, A_1\} = \boldsymbol{H}_{1,0}^{\intercal}\boldsymbol{\beta}_{1,0}^k + A_1\boldsymbol{H}_{1,1}^{\intercal}\boldsymbol{\beta}_{1,1}^k$. Exploratory analyses reveal little evidence of heteroskedasticity at the first-stage. Thus, we opt to estimate a constant residual variance for both terms following the mean modeling steps. After the mean and variance modeling steps, we use a Gaussian copula to estimate the joint conditional distribution of the standardized residuals of $\{m(\boldsymbol{H}_2), c(\boldsymbol{H}_2)\}$ given $\boldsymbol{H}_1$ and A_1, resulting in our estimate of $G(\cdot, \cdot \mid \boldsymbol{h}_1, a_1)$ which we denote by $\widehat{G}(\cdot, \cdot \mid \boldsymbol{h}_1, a_1)$.

The estimated first-stage optimal rule is $\widehat{\pi}^{\text{TIQ}}_{1,\lambda}(\boldsymbol{h}_1) = \arg\min_{a_1} \int \widehat{F}_\epsilon(21-u-|v|)d\widehat{G}(u,v \mid \boldsymbol{h}_1, a_1)$. At stage two, $\widehat{\pi}^*_2(\boldsymbol{h}_2) = \text{sgn}(-1.66+0.15*\texttt{qids1}-4.03*\texttt{slope1}-0.68*\texttt{A1})$ is the estimated optimal treatment. Based on Remark 1 in Section 2.1, we compare the estimated first-stage treatment recommendations to those recommended by the mean-optimal rule, $\arg\min_{a_1} \int(-u-|v|)d\widehat{G}(u,v \mid \boldsymbol{h}_1, a_1)$, for each observed $\boldsymbol{h}_1$ in the data. Only one patient out of 132 is recommended differently. In addition, the difference in raw values of $\int \widehat{F}_\epsilon(21-u-|v|)d\widehat{G}(u,v \mid \boldsymbol{h}_1, a_1)$ for $a_1 = 1, -1$ as well as $\int(-u-|v|)d\widehat{G}(u,v \mid \boldsymbol{h}_1, a_1)$ for $a_1 = 1, -1$ are the smallest for this particular patient. Thus, the treatment discrepancy is most likely due to a near-zero treatment effect for this patient.

We compare TIQ-learning to the Q-learning analysis of Schulte et al. (2012) and binary Q-learning (Chakraborty and Moodie, 2013). Comparing the results to Q-learning, which maximizes the expected value of Y, supports the claim that TIQ-learning and mean optimization are equivalent for this subset of the STAR*D data. The first step of Q-learning is to model the conditional expectation of Y given $\boldsymbol{H}_2$ and A_2 which is the same as the first step of TIQ-learning. Thus, we use the same model and estimated decision rule at stage two given in Step 1 of the TIQ-learning algorithm. Next, we model the conditional expectation of $\widetilde{Y} = \boldsymbol{H}^\intercal_{2,0}\boldsymbol{\beta}_{2,0} + |\boldsymbol{H}^\intercal_{2,1}\boldsymbol{\beta}_{2,1}|$, where $\widetilde{Y}$ is the predicted future optimal outcome at stage one. We specify the working model $E(\widetilde{Y} \mid \boldsymbol{H}_1, A_1) = \boldsymbol{H}^\intercal_{1,0}\boldsymbol{\beta}^Q_{1,0} + A_1\boldsymbol{H}^\intercal_{1,1}\boldsymbol{\beta}^Q_{1,1}$, where $\boldsymbol{H}^\intercal_{1,0} = \boldsymbol{H}^\intercal_{1,1} = (1, \boldsymbol{X}^\intercal_1)^\intercal$ and $\boldsymbol{X}_1 = (\texttt{qids0}, \ \texttt{slope0})^\intercal$. We fit the model using least squares. Then, the Q-learning estimated optimal first-stage rule is $\widehat{\pi}^{\text{Q}}_{1,\lambda}(\boldsymbol{h}_1) = \text{sgn}(-0.95 + 0.13 * \texttt{qids1} + 2.17 * \texttt{slope1})$. Q-learning recommends treatment differently at the first stage for only one of the 132 patients in the data. In addition, the estimated value of the TIQ- and Q-learning regimes are nearly the same and are displayed in Table 2. Binary Q-learning recommends treatment differently than TIQ-learning for 18 patients at the first stage, and the estimated value of the binary Q-learning regime is slightly lower than TIQ- and Q-learning. Also included in Table 2 are value estimates for four non-dynamic regimes that treat everyone according to the decision rules $\pi_1(\boldsymbol{h}_1) = a_1$ and $\pi_2(\boldsymbol{h}_2) = a_2$ for $a_1 \in \{-1, 1\}$ and $a_2 \in \{-1, 1\}$. We estimate these values using the Augmented Inverse Probability Weighted Estimator given in Zhang

et al. (2013).

In summary, it appears that TIQ-learning and Q-learning perform similarly for this subset of the STAR*D data. This may be due to the lack of heteroskedasticity at the first stage. Thus, maximizing the end-of-stage-two QIDS using mean-optimal techniques seems appropriate and, in practice, equivalent to maximizing remission probabilities for each patient with TIQ-learning.

6. DISCUSSION

We have proposed modeling frameworks for estimating optimal dynamic treatment regimes in settings where a non-mean distributional summary is the intended outcome to optimize. Threshold Interactive Q-learning (TIQ-learning) estimates a regime that maximizes the mass of the response distribution that exceeds a constant or patient-dependent threshold. Although TIQ-learning is just a special case of optimizing regimes based on binary responses, it is a helpful precursor for the development of Quantile Interactive Q-learning (QIQ-learning), which maximizes a prespecified quantile of the response distribution. If the focus is simply on optimizing a dynamic treatment regime for a binary threshold outcome, rather than quantile optimization, recent nonparametric techniques tailored to the binary outcome setting may be preferred to avoid problems due to model misspecification (Zhao et al., 2012, 2015). For example, it is possible to prespecify a class of regimes and directly optimize within that class by utilizing the value function,

$$V^{\pi_1,\pi_2} = \frac{\sum_{i=1}^{n} \mathbb{1}_{Y_i>\lambda} \mathbb{1}_{A_{1,i}=\pi_1(\boldsymbol{H}_{1,i})} \mathbb{1}_{A_{2,i}=\pi_2(\boldsymbol{H}_{2,i})}}{\sum_{i=1}^{n} \mathbb{1}_{A_{1,i}=\pi_1(\boldsymbol{H}_{1,i})} \mathbb{1}_{A_{2,i}=\pi_2(\boldsymbol{H}_{2,i})}}.$$

To our knowledge, this is the first attempt to estimate an optimal treatment regime that targets a quantile. Despite generalizations presented in the supplementary material, some researchers might be wary of the assumptions needed for consistency of the estimated optimal regime or the requisite modeling of the main effect function. An interesting future direction would be to develop a flexible framework for QIQ-learning that depends only on models for the treatment contrast function and not on the main effect term. In addition, it has been shown that Q-learning, A-learning,

and g-estimation lead to identical estimators in certain cases but that Q-learning is less efficient (Chakraborty et al., 2010; Schulte et al., 2012). It is possible a g-estimation approach exists for maximizing quantiles based on structural nested distribution models (Robins, 2000; Vansteelandt et al., 2014).

Our proposed methods are designed for the two-stage setting, this is an important development given that many completed and ongoing SMART studies have this structure (Projects Using SMART, 2012; Laber, 2013). Here we considered binary treatments at both stages. In principle, the proposed methods can be extended to settings with more than two treatments at each stage by modeling additional treatment contrasts. Formalization of this idea merits further research.

7. SUPPLEMENTARY MATERIALS

Online supplementary materials include discussions of modeling adjustments for heteroskedastic second-stage errors and patient-specific thresholds, a proof of Lemma 3.1 and toy example illustrating where this lemma does not apply, additional simulation results, and proofs of the theorems in Section 3.3.

REFERENCES

Bembom, O. and van der Laan, M. J. (2008). Analyzing sequentially randomized trials based on causal effect models for realistic individualized treatment rules. *Statistics in medicine*, 27(19):3689–3716.

Berkowitz, R. I., Wadden, T. A., Gehrman, C. A., Bishop-Gilyard, C. T., Moore, R. H., Womble, L. G., Cronquist, J. L., Trumpikas, N. L., Katz, L. E. L., and Xanthopoulos, M. S. (2010). Meal Replacements in the Treatment of Adolescent Obesity: A Randomized Controlled Trial. *Obesity*, 19(6):1193–1199.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.

Chakraborty, B. and Moodie, E. E. (2013). *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*, volume 76. Springer Science & Business Media.

Chakraborty, B., Murphy, S. A., and Strecher, V. J. (2010). Inference for Non-Regular Parameters in Optimal Dynamic Treatment Regimes. *Statistical Methods in Medical Research*, 19(3):317–343.

Fava, M., Rush, A. J., Trivedi, M. H., Nierenberg, A. A., Thase, M. E., Sackeim, H. A., Quitkin, F. M., Wisniewski, S. R., Lavori, P. W., Rosenbaum, J. F., Kupfer, D. J., and STAR*D Investigators Group (2003). Background and Rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Study. *Psychiatric Clinics of North America*, 26(2):457 – 94.

Glick, N. (1974). Consistency Conditions for Probability Estimators and Integrals of Density Estimators. *Utilitas Mathematica*, 6:61–74.

Goldberg, Y. and Kosorok, M. R. (2012). Q-learning with Censored Data. *Annals of statistics*, 40(1):529.

Goldberg, Y., Song, R., and Kosorok, M. R. (2013). Adaptive *Q*-learning. *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner*, pages 150 – 62.

Henderson, R., Ansell, P., and Alshibani, D. (2010). Regret-regression for Optimal Dynamic Treatment Regimes. *Biometrics*, 66(4):1192–1201.

Huang, X. and Ning, J. (2012). Analysis of multi-stage treatments for recurrent diseases. *Statistics in medicine*, 31(24):2805–2821.

Huang, X., Ning, J., and Wahed, A. S. (2014). Optimization of individualized dynamic treatment regimes for recurrent diseases. *Statistics in medicine*, 33(14):2363–2378.

Kang, C., Janes, H., and Huang, Y. (2014). Combining biomarkers to optimize patient treatment recommendations. *Biometrics*, 70(3):695–707.

Laber, E. B. (2013). Example SMARTs. http://www4.stat.ncsu.edu/ laber/smart.html.

Laber, E. B., Linn, K. A., and Stefanski, L. A. (2014a). Interactive model building for q-learning. *Biometrika*, page asu043.

Laber, E. B., Lizotte, D. J., and Ferguson, B. (2014b). Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*, 70(1):53–61.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014c). Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, 8(1):1225.

Lavori, P. W. and Dawson, R. (2000). A Design for Testing Clinical Strategies: Biased Adaptive Within-Subject Randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):29–38.

Lavori, P. W. and Dawson, R. (2004). Dynamic Treatment Regimes: Practical Design Considerations. *Clinical Trials*, 1(1):9–20.

Linn, K., Laber, E., and Stefanski, L. (2015). Constrained estimation for competing outcomes. In *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*, pages 1–23. CRC Press.

Luedtke, A. R. and van der Laan, M. J. (2015). Optimal dynamic treatments in resource-limited settings.

McKeague, I. W. and Qian, M. (2013). Evaluation of Treatment Policies Based on Functional Predictors. *Statistica Sinica*, submitted.

Moodie, E. E., Dean, N., and Sun, Y. R. (2014). Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences*, 6(2):223–243.

Moodie, E. E. M. and Richardson, T. S. (2010). Estimating Optimal Dynamic Regimes: Correcting Bias Under the Null. *Scandinavian Journal of Statistics*, 37(1):126 – 146.

Murphy, S. A. (2003). Optimal Dynamic Treatment Regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Murphy, S. A. (2005a). A Generalization Error for Q-learning. *Journal of Machine Learning Research*, 6(7):1073 – 1097.

Murphy, S. A. (2005b). An Experimental Design for the Development of Adaptive Treatment Strategies. *Statistics in Medicine*, 24(10):1455–1481.

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. *The International Journal of Biostatistics*, 6(2).

Petersen, M., Schwab, J., Gruber, S., Blaser, N., Schomaker, M., and van der Laan, M. (2014). Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of Causal Inference*, 2(2):147–185.

Projects Using SMART (2012). The Methodology Center at Pennsylvania State University. http://methodology.psu.edu/ra/adap-inter/projects.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.

Robins, J., Orellana, L., and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27(23):4678–4721.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer.

Robins, J. M. (2004). Optimal Structural Nested Models for Optimal Sequential Decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, pages 189–326. Springer New York.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of educational Psychology*, 66(5):688.

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., Thase, M. E., Nierenberg, A. A., Quitkin, F. M., Kashner, T., Kupfer, D. J., Rosenbaum, J. F., Alpert, J., Stewart, J. W., McGrath, P. J., Biggs, M. M., Shores-Wilson, K., Lebowitz, B. D., Ritz, L., Niederehe, G., and STAR*D Investigators Group (2004). Sequenced Treatment Alternatives to Relieve Depression (STAR*D): Rationale and Design. *Controlled Clinical Trials*, 25(1):119 – 142.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). Q- and A-learning Methods for Estimating Optimal Dynamic Treatment Regimes. *arXiv:1202.4177 [stat.ME]*.

Shortreed, S. M., Laber, E., Stroup, T. S., Pineau, J., and Murphy, S. A. (2014). A Multiple Imputation Strategy for Sequential Multiple Assignment Randomized Trials. *Statistics in Medicine*, to appear.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC.

Song, R., Wang, W., Zeng, D., and Kosorok, M. R. (2015). Penalized Q-Learning for Dynamic Treatment Regimes. *Statistica Sinica*, 25:901–920.

van der Laan, M. J. (2006). Causal effect models for intention to treat and realistic individualized treatment rules.

van der Laan, M. J. and Petersen, M. L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1).

van der Laan, M. J., Petersen, M. L., and Joffe, M. M. (2005). History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *The International Journal of Biostatistics*, 1(1).

Vansteelandt, S., Joffe, M., et al. (2014). Structural nested models and g-estimation: The partially realized promise. *Statistical Science*, 29(4):707–731.

Watkins, C. J. C. H. (1989). Learning from Delayed Rewards. *PhD Thesis, University of Cambridge, England*.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-Learning. *Machine Learning*, 8:279–292.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012a). Estimating Optimal Treatment Regimes from a Classification Perspective. *Stat*, 1(1):103–114.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012b). A Robust Method for Estimating Optimal Treatment Regimes. *Biometrics*, 68(4):1010–1018.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust Estimation of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions. *Biometrika*.

Zhang, Z., Chen, Z., Troendle, J. F., and Zhang, J. (2012c). Causal Inference on Quantiles with an Obstetric Application. *Biometrics*, 68(3):697–706.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 0(0):00–00.

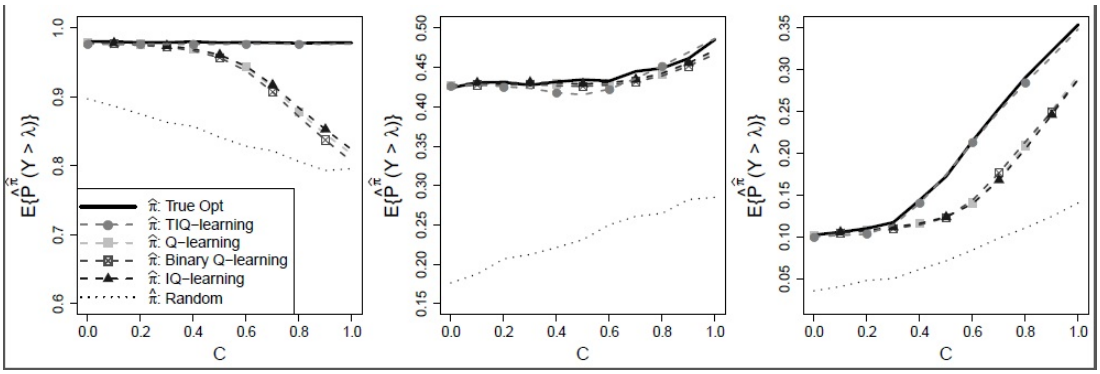

Figure 1: *Left to Right:* $\lambda = -2, 2, 4$. Solid black, true optimal threshold probabilities; dotted black, probabilites under randomization; dashed with circles/squares/crossed squares/triangles, probabilities under TIQ-, Q-, binary Q-, and Interactive Q-learning, respectively.

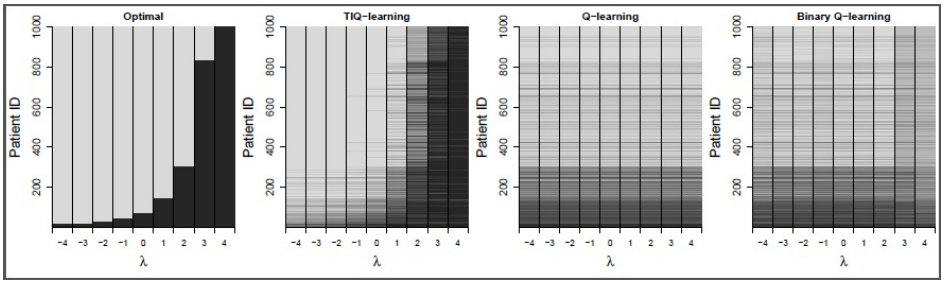

Figure 2: *From left:* True optimal first-stage treatments for 1,000 test set patients when $\lambda = -4, -3, ..., 4$, coded light gray when $\pi^{\text{TIQ}}_{1,\lambda}(h_1) = 1$ and dark gray otherwise; TIQ-learning estimated optimal first-stage treatments; Q-learning estimated optimal first-stage treatments, plotted constant in λ to aid visual comparison; and binary Q-learning estimated optimal first-stage treatments for each λ.

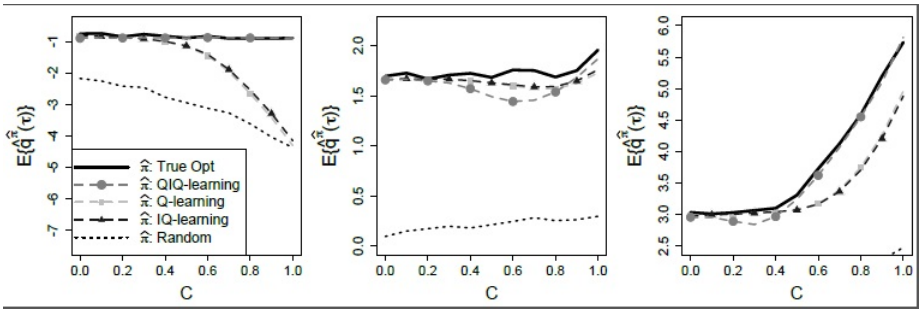

Figure 3: *Left to Right:* $\tau = 0.1, 0.5, 0.75$. Solid black, true optimal quantiles; dotten black, quantiles under randomization; dashed with circles/squares/triangles, quantiles under QIQ-, *Q*-, and IQ-learning, respectively.

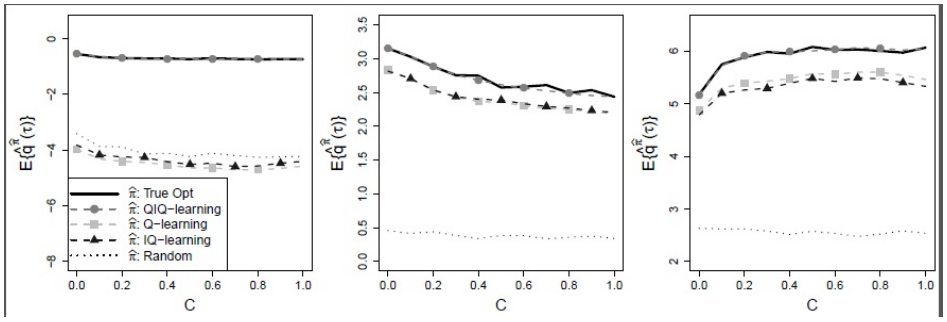

Figure 4: *Left to Right:* $\tau = 0.1, 0.5, 0.75$. Solid black, true optimal threshold probabilities; dotted black, probabilites under randomization; dashed with circles/squares/triangles, probabilities under TIQ-, Q-, and Interactive Q-learning, respectively. Training set size of $n = 500$.

Table 1: Variables used in the STAR*D analysis.

Variable	Description
`qids0`	mean QIDS during the baseline stage.
`slope0`	pre-randomization QIDS improvement; the difference between the final and initial baseline-stage QIDS scores, divided by time spent in the baseline stage.
`qids1`	mean stage-one QIDS.
`slope1`	first-stage QIDS improvement; the difference between the final and initial first-stage QIDS scores, divided by time spent in the first randomized stage.
`A1`	First-stage treatment; 1="SSRI" and -1="non-SSRI."
`A2`	Second-stage treatment; 1="NTP" for Nortriptyline and -1="MIRT" for Mirtazapine.
`Y`	27 minus final QIDS score, measured at the end of stage two.

Table 2: Estimated value of dynamic and non-dynamic regimes using the Adaptive Inverse Probability Weighted Estimator.

	Estimated Value
TIQ-learning	0.24
Q-learning	0.23
Binary Q-learning	0.19
(1, 1)	0.13
(-1, 1)	0.24
(1, -1)	0.07
(-1, -1)	0.12