Published in final edited form as:
Biometrika. 2014 October 20; 101(4): 831–847. doi:10.1093/biomet/asu043.

Interactive model building for *Q*-learning

Eric B. Laber, **Kristin A. Linn**, and **Leonard A. Stefanski**
Department of Statistics, North Carolina State University, 2311 Stinson Drive, 5216 SAS Hall, Raleigh, North Carolina, 27695-8203, USA

Eric B. Laber: laber@stat.ncsu.edu; Kristin A. Linn: kalinn@ncsu.edu; Leonard A. Stefanski: stefansk@ncsu.edu

Summary

Evidence-based rules for optimal treatment allocation are key components in the quest for efficient, effective health care delivery. Q-learning, an approximate dynamic programming algorithm, is a popular method for estimating optimal sequential decision rules from data. Q-learning requires the modeling of nonsmooth, nonmonotone transformations of the data, complicating the search for adequately expressive, yet parsimonious, statistical models. The default Q-learning working model is multiple linear regression, which is not only provably misspecified under most data-generating models, but also results in nonregular regression estimators, complicating inference. We propose an alternative strategy for estimating optimal sequential decision rules for which the requisite statistical modeling does not depend on nonsmooth, nonmonotone transformed data, does not result in nonregular regression estimators, is consistent under a broader array of data-generation models than Q-learning, results in estimated sequential decision rules that have better sampling properties, and is amenable to established statistical approaches for exploratory data analysis, model building, and validation. We derive the new method, IQ-learning, via an interchange in the order of certain steps in Q-learning. In simulated experiments IQ-learning improves on Q-learning in terms of integrated mean squared error and power. The method is illustrated using data from a study of major depressive disorder.

Some key words

Dynamic Treatment Regime; Personalized Medicine; Treatment Selection

1. Introduction

Clinical treatment decisions are based on a patient's treatment history and current health status. Sequential decision rules, also known as dynamic treatment regimes, formalize this process by specifying a sequence of decision rules, one for each treatment decision, that take as input a patient's treatment and covariate history and output recommended treatments. An optimal sequential decision rule is one that maximizes a desirable clinical outcome. The sequential nature of clinical decision making problems has led researchers to estimate

Supplementary Material
Supplementary material available at Biometrika online includes proofs of the theoretical results stated in Section 2, additional simulation results, and a list of covariates used in Section 4.

optimal sequential decision rules using approximate dynamic programming procedures (Robins, 2004; Murphy, 2005a). Q-learning with function approximation, hereafter Q-learning, is one such popular method (Murphy, 2005a). However, Q-learning and similar variants of it involve modeling nonsmooth, nonmonotone functions of the data. Nonmonotonicity complicates the regression function whereas nonsmoothness imparts nonregularity to estimators. Inference in the presence of such nonregularity has been well-studied by Robins (2004), Chakraborty et al. (2010), Laber et al. (2014), and Moodie and Richardson (2010). However, much less attention has been directed toward the effect of nonsmoothness on the equally-important applied problems of model building and diagnostics. The preceding references all study linear working models, making little mention of their appropriateness or how to interactively build a model using data. As we demonstrate, even under simple generative models, linear working models result in questionable fits.

Rather than develop specialized exploratory and model building techniques for Q-learning, we propose to model the data before applying the necessary nonmonotone, nonsmooth operations. Because standard interactive model building techniques can be used with our new version of Q-learning, we call it IQ-learning for interactive Q-learning. Interactive model building is an essential part of extracting meaningful information from data (Henderson and Velleman, 1981; Cook and Weisberg, 1982; Henderson et al., 2010; Rich et al., 2010; Chakraborty and Moodie, 2013). This is especially true when the analysis is intended to inform clinical practice or provide scientific insight, and thus IQ-learning is especially attractive in applications.

IQ-learning has several advantages over Q-learning. For a large class of generative models, IQ-learning involves only simple, well-studied, and well-understood conditional mean and variance modeling of smooth transformations of the data, resulting in better fitting and more interpretable models (Carroll and Ruppert, 1988). Furthermore, inference for coefficients indexing the working models in IQ-learning is greatly simplified by regular normal limit theory. However, IQ-learning does not resolve the problem of nonregularity of the estimated non-terminal Q-functions, just of the coefficients on which they depend. Nevertheless, this is an important distinction from Q-learning and related methods. This issue is discussed further in Section 2.3.

2. Q- and IQ-Learning

2·1. Setup

For simplicity we consider the case of two treatment decisions and two treatment options per stage. In addition to being the most common in practice (see www4.stat.ncsu.edu/~laber/smarts.html), the two-stage, binary-treatment setting facilitates exposition while maintaining the key features of more general problems. While we focus on randomized studies, the proposed approach is valid for observational data under the same causal conditions required for Q-learning (Schulte et al., 2012).

Training data for the two-stage, two-treatment case, $\mathscr{D}=\{(X_{1,i}, A_{1,i}, X_{2,i}, A_{2,i}, Y_i)\}_{i=1}^{n}$, consists of independent identically distributed copies of the quintuple (X_1, A_1, X_2, A_2, Y)

containing data collected on a single subject. Each quintuple is time ordered and thus called a trajectory: $X_1 \in \mathbb{R}^{p1}$ is baseline covariate information; $A_1 \in \{-1, 1\}$ is the first treatment; $X_2 \in \mathbb{R}^{p2}$ is covariate information collected between the first and second treatment assignments; $A_2 \in \{-1, 1\}$ is the second treatment; and $Y \in \mathbb{R}$ is the outcome, coded so that higher values coincide with more desirable clinical outcomes. For notational compactness and conformity with established practice, we denote the information available prior to the t^{th} treatment assignment by H_t. Thus, $H_1 = X_1$ and $H_2=(X_1^\top, A_1, X_2^\top)^\top$.

2·2. Q-Learning

The goal of Q- and IQ-learning is estimation of a pair of decision rules $\pi = (\pi_1, \pi_2)$ such that π_t maps the domain of H_t into the set of treatments, $\pi_t : \mathscr{H}_t \mapsto \{-1, 1\}$. An optimal sequential decision rule π^{opt} maximizes expected outcome. Let E^π denote expectation under the restriction $A_t = \pi_t(H_t)$; then π^{opt} satisfies $E^{\pi^{\text{opt}}}(Y) = \sup_\pi E^\pi(Y)$.

Q-learning, a regression-based, approximate dynamic programming algorithm, is commonly used to estimate π^{opt} (Murphy, 2005a). The algorithm depends on the Q-functions,

$$Q_2(h_2, a_2)=E(Y|H_2=h_2, A_2=a_2), \quad (1)$$

$$Q_1(h_1, a_1)=E\left\{\max_{a_2\in\{-1,1\}} Q_2(H_2, a_2)|H_1=h_1, A_1=a_1\right\}. \quad (2)$$

Thus, $Q_t(h_t, a_t)$ measures the quality of treatment a_t when assigned to a patient with history h_t, assuming that optimal treatment decisions are made in future stages (Sutton and Barto, 1998). If the Q-functions were known, the optimal sequential decision rule could be determined using dynamic programming (Bellman, 1957), yielding the solution $\pi_t^{\text{opt}}(h_t)=\arg\max_{a_t\in\{-1,1\}} Q_t(h_t, a_t)(t=1, 2)$. The Q-functions are seldom known, and in practice Q-learning mimics the dynamic programming solution by replacing the unknown conditional expectations with fitted regression models. For reasons of data scarcity, model parsimony, and simplicity, it is common to use linear models for the Q-functions:

$$Q_t(h_t, a_t;\beta_t)=h_{t,0}^\top\beta_{t,0}+a_t h_{t,1}^\top\beta_{t,1}, t=1, 2, \quad (3)$$

where $h_{t,0}$ and $h_{t,1}$ are, possibly the same, subvectors of h_t, and $\beta_t=(\beta_{t,0}^\top, \beta_{t,1}^\top)^\top$. Q-learning with linear models consists of three steps:

Q1 estimate β_2, and hence Q_2, via least squares regression of Y on H_2 and A_2 using the working model (3), i.e., $\hat{\beta}_2=\arg\min_{\beta_2}\sum_{i=1}^n\{Y_i - Q_2(H_{2,i}, A_{2,i};\beta_2)\}^2$;

Q2 calculate predicted future outcomes $\tilde{Y}$ assuming optimal Stage 2 decisions

$$\tilde{Y}=\max_{a_2\in\{-1,1\}} Q_2(H_2, a_2;\hat{\beta}_2)=H_{2,0}^\top\hat{\beta}_{2,0}+|H_{2,1}^\top\hat{\beta}_{2,1}|;$$

then estimate β_1, and hence Q_1, via least squares regression of $\tilde{Y}$ on H_1 and A_1 using the working model (3), i.e.,

$$\hat{\beta}_1 = \arg \min_{\beta_1} \sum_{i=1}^{n} \left\{ \tilde{Y}_i - Q_1(H_{1,i}, A_{1,i}; \beta_1) \right\}^2 ; \text{ and}$$

Q3 calculate the estimated Q-learning optimal treatment policy $\hat{\pi}^Q = (\hat{\pi}_1^Q, \hat{\pi}_2^Q)$ as

$$\hat{\pi}_t^Q(h_t) = \arg \max_{a_t \in \{-1,1\}} Q_t(h_t, a_t; \hat{\beta}_t).$$

There is nothing unusual about the regression modeling in the first step, which is amenable to well-studied techniques and diagnostics. The same is not true of the modeling in the second step. The absolute value function in the definition of $\tilde{Y}$ means that even if the relationships among the components of H_1 and H_2 are approximately linear, the dependence of $\tilde{Y}$ on H_1 would necessarily be nonmonotone. Thus the working model for $\tilde{Y}$ commonly used in practice is generally wrong. Furthermore, there are no simple transformations of the observed variables or simple modifications to the working model that render the true regression of $\tilde{Y}$ on H_1 and A_1 linear, even asymptotically. We illustrate this point with data generated from the model

$$\begin{array}{ll} X_1 \sim \mathrm{N}(0, \sigma^2), & \xi \sim \mathrm{N}(0, \tau^2), X_2 = \zeta X_1 + \xi, \\ A_t \sim \mathrm{Unif}\{-1, 1\}, t = 1, 2, & \phi \sim \mathrm{N}(0, \gamma^2), Y = 1.25 A_1 A_2 + A_2 X_2 - A_1 X_1 + \phi, \end{array} \quad (4)$$

where σ^2, τ^2, ζ, and γ^2 are fixed parameters. In most applications one expects $\zeta = \mathrm{cov}(X_1, X_2) \neq 0$. Treatments are randomly assigned at each stage as in a sequential multiple assignment randomized trial design (Murphy, 2005b; see also Lavori and Dawson, 2000, 2004). The working model in (3) for $Q_2(H_2, A_2)$ is correct, so the resulting predicted value $\tilde{Y}$ approximates the true fitted value $\tilde{Y}_{\text{True}} = |1.25A_1 + X_2| - A_1X_1$. It follows that the regression in Step Q2 approximates the regression of $\tilde{Y}_{\text{True}}$ on $H_1 = X_1$ and A_1. Substituting for X_2 in $\tilde{Y}$ shows that $\tilde{Y}_{\text{True}} = |1.25A_1 + \zeta X_1 + \xi| - A_1X_1$, from which it is apparent that $E(\tilde{Y}_{\text{True}}|X_1, A_1)$ is linear in X_1 for fixed A_1 only in the unlikely case that $\zeta = 0$. Thus correlation between X_1 and X_2 induces a nonlinear dependence of $\tilde{Y}$ on X_1.

The left-hand panel of Figure 1 displays a scatterplot of $\tilde{Y}$ against X_1 for each value of A_1 based on 1,000 random draws from model (4) with $\zeta = 0.85$, $\sigma = 1$, and $\tau = \gamma = 1/\sqrt{2}$, using the Q-learning algorithm to calculate $\tilde{Y}$. The figure illustrates nonlinearity in the regression of $\tilde{Y}$ on X_1 and also heteroscedastic variation induced by the max operation in Step Q2. As this toy model makes clear, identifying the correct form of the regression model for $E(\tilde{Y}_{\text{True}} | H_1, A_1)$ and fitting it efficiently would be difficult in the realistic case that the data-generating model is unknown; one approach would be to adopt nonparametric models for the Q-functions (e.g., Zhao et al., 2011; Moodie et al., 2013), but some clinicians are wary of black-box approaches and it 125 can be difficult to glean scientific knowledge from these models. Thus, common practice is to ignore the problem and settle for the best approximation afforded by fitting linear models. This problem is shared by variants of Q-learning (e.g., A-learning, Murphy, 2003; Blatt et al., 2004; Robins, 2004; Schulte et al., 2012). In contrast, the right-hand panel of Fig. 1 shows the first-stage regression model that

must be fit in our proposed method, IQ-learning, described next. It is a common analysis of covariance model in X_1 and A_1.

2·3. IQ-Learning

IQ-learning replaces the difficult problem of modeling the predicted future optimal outcomes $\tilde{Y}=\max_{a_2\in\{-1,1\}}Q_2(H_2,a_2;\hat{\beta}_2)=H_{2,0}^{\top}\hat{\beta}_{2,0}+|H_{2,1}^{\top}\hat{\beta}_{2,1}|$ with two ordinary mean-variance function modeling problems. Its practical advantages result from the fact that there is a wealth of models and theory for mean-variance function modeling (Carroll and Ruppert, 1988). Thus it has the potential for better model building and diagnostics. The modeling required is familiar and is generally interactive. We first describe the IQ-learning algorithm in general terms and then discuss special cases that are useful in practice.

Whereas Q-learning models $\max_{a2\in\{-1,1\}} Q_2(H_2, a_2)$ directly, IQ-learning starts with the Q_2 contrast and main-effect functions: $\Delta(H_2) = \{Q_2(H_2, 1) - Q_2(H_2, -1)\}/2$; $\mu(H_2) = \{Q_2(H_2, 1) + Q_2(H_2, -1)\}/2$. The contrast and main-effect functions are linear, and hence smooth and monotone functions of $Q_2(\cdot)$. Let $g_{h1,\,a1}(\cdot)$ denote the conditional distribution of the contrast $\Delta(H_2)$ given $H_1 = h_1$ and $A_1 = a_1$. With these definitions $Q_1(h_1, a_1)$ defined in (2) can be written as

$$Q_1(h_1,a_1)=E\{\mu(H_2)|H_1=h_1,A_1=a_1\}+\int|z|g_{h_1,a_1}(z)dz. \quad (5)$$

The IQ-learning estimator of $Q_1(h_1, a_1)$ has the form

$$\hat{Q}_1^{\mathrm{IQ}}(h_1,a_1)=\hat{L}(h_1,a_1)+\int|z|\hat{g}_{h_1,a_1}(z)dz, \quad (6)$$

where $\hat{L}(h_1, a_1)$ and $\hat{g}_{h1,\,a1}(\cdot)$ are estimators of $E\{\mu(H_2) \mid H_1 = h_1, A_1 = a_1\}$ and $g_{h1,\,a1}(\cdot)$.

Let $\hat{Q}_2(H_2, A_2)$ denote the estimator obtained in Step Q1 of the Q-learning algorithm. Define the estimated main-effect and contrast functions

$$\hat{\mu}(H_2)=\frac{1}{2}\left\{\hat{Q}_2(H_2,1)+\hat{Q}_2(H_2,-1)\right\},\ \hat{\Delta}(H_2)=\frac{1}{2}\left\{\hat{Q}_2(H_2,1)-\hat{Q}_2(H_2,-1)\right\}. \quad (7)$$

Then $\hat{L}(h_1, a_1)$ is obtained by modeling the regression of $\hat{\mu}(H_2)$ on H_1 and A_1 for which linear models are often adequate as no unusual transformations are involved. Obtaining $\hat{g}_{h1,\,a1}(\cdot)$ is accomplished by estimating the conditional distribution of $\hat{\Delta}(H_2)$ given H_1 and A_1. For this we exploit mean-variance function modeling as explained in Section 2·4. Thus we have the following algorithm for IQ-learning:

IQ1 use Step Q1 of the Q-learning algorithm to obtain $\hat{\beta}_2$ and $\hat{Q}_2^{\mathrm{IQ}}(H_2,A_2)=Q_2(H_2,A_2;\hat{\beta}_2)$;

IQ2 **a.** regress the estimated main-effect function $\hat{\mu}(H_2)$ from (7) on H_1 and A_1 to obtain an estimator $\hat{L}(h_1, a_1)$ of $E\{\mu(H_2) \mid H_1 = h_1, a_1\}$;

b. model the conditional distribution of the estimated contrast function $\hat{\Delta}(H_2)$ from (7) given $H_1 = h_1$ and $A_1 = a_1$ to obtain an estimator $\hat{g}_{h_1, a_1}(\cdot)$ of $g_{h_1, a_1}(\cdot)$;

c. combine the estimators from IQ2a and IQ2b to obtain

$$\hat{Q}_1^{\mathrm{IQ}}(h_1, a_1) = \hat{L}(h_1, a_1) + \int |z| \hat{g}_{h_1, a_1}(z) dz;$$

IQ3 define the IQ-learning estimated optimal treatment policy $\hat{\pi}^{\mathrm{IQ}} = (\hat{\pi}_1^{\mathrm{IQ}}, \hat{\pi}_2^{\mathrm{IQ}})$ so that $\hat{\pi}_t^{\mathrm{IQ}}(h_t) = \arg \max_{a_t \in \{-1,1\}} \hat{Q}_t^{\mathrm{IQ}}(h_t, a_t)$.

Completing our algorithm requires specific models for Steps IQ1, IQ2a, and IQ2b. As noted previously, Steps IQ1 and IQ2a are usually straightforward and linear models will often suffice. We now show how to accomplish the modeling in Step IQ2b efficiently and with sufficient flexibility for many applications by using mean-variance models.

2·4. Location-Scale Working Models for $g_{h_1, a_1}(\cdot)$

Henceforth we consider mean-variance, location-scale estimators of $g_{h_1, a_1}(\cdot)$ of the form

$$\hat{g}_{h_1, a_1}(z) = \frac{1}{\hat{\sigma}(h_1, a_1)} \hat{\phi} \left\{ \frac{z - \hat{m}(h_1, a_1)}{\hat{\sigma}(h_1, a_1)} \right\}, \quad (8)$$

where $\hat{m}(h_1, a_1)$ is an estimator of $m(h_1, a_1) = E\{\Delta(H_2) \mid H_1 = h_1, A_1 = a_1\}$, $\hat{\sigma^2}(h_1, a_1)$ is an estimator of $\sigma^2(h_1, a_1) = E[\{\Delta(H_2) - m(h_1, a_1)\}^2 \mid H_1 = h_1, A_1 = a_1]$, and $\hat{\varphi}$ is an estimator of the density of the standardized residuals $\{\Delta(H_2) - m(h_1, a_1)\} / \sigma(h_1, a_1)$, say φ_{h_1, a_1}, which we assume does not depend on the history h_1 or the treatment a_1. In other words, we assume that all of the dependence of $\Delta(H_2)$ on (H_1, A_1) is captured by the conditional mean and variance functions. The great success of mean-variance function modeling (Carroll and Ruppert, 1988) suggests that this assumption is reasonable quite generally; however, we also note that substantial departures from the assumption can be investigated by stratifying on h_1 and a_1 and comparing higher-order moments, such as skewness and kurtosis, or nonparametric density estimates of the empirical residuals $\{\hat{\Delta}(H_2) - \hat{m}(H_1, A_1)\} / \hat{\sigma}(H_1, A_1)$ across the strata. We now describe two special cases of the estimator in (8).

Let φ denote the density of a standard normal random variable. A simple but useful estimator of $g_{h_1, a_1}(\cdot)$ is the normal location-scale model:

$$\hat{g}_{h_1, a_1}^N(z) = \frac{1}{\hat{\sigma}(h_1, a_1)} \phi \left\{ \frac{z - \hat{m}(h_1, a_1)}{\hat{\sigma}(h_1, a_1)} \right\}, \quad (9)$$

which is a special case of (8) with $\hat{\varphi} = \varphi$. An advantage of this model is that $\int |z| \hat{g}_{h_1, a_1}^N(z) dz$ can be evaluated in closed form. In particular,

$$\int |z| \hat{g}_{h_1, a_1}^N(z) dz = \hat{m}(h_1, a_1) \left[2\Phi \left\{ \frac{\hat{m}(h_1, a_1)}{\hat{\sigma}(h_1, a_1)} \right\} - 1 \right] + 2\hat{\sigma}(h_1, a_1) \phi \left\{ \frac{\hat{m}(h_1, a_1)}{\hat{\sigma}(h_1, a_1)} \right\}, \quad (10)$$

where Φ is the standard normal cumulative distribution function. If the mean and variance functions are correctly specified and $\varphi_{h_1, a_1} = \varphi$, then the IQ-learning location-scale model is correct. As commonly implemented, Q-learning fits a misspecified model and thus estimators are not consistent. The closed form expression in (10) makes it possible to study the bias in Q-learning when the true data-generating model is a normal mean-variance function model. Details are in the Supplementary Material.

The normal location-scale model assumes that $\varphi_{h_1, a_1} = \varphi$, the standard normal density. Violations of normality can be diagnosed via examination of the observed standardized residuals, $\hat{e}_i = \{\hat{\Delta}(H_{2,i}) - m\{\hat{}(H_{1,i}, A_{1,i})\}/\sigma\hat{}(H_{1,i}, A_{1,i})$ ($i = 1, \ldots, n$). When greater modeling flexibility is desired, the normality assumption can be dropped and the empirical distribution of the $\hat{e}_i$ used instead. Defining

$$\hat{G}_{h_1,a_1}(z)=\frac{1}{n}\sum_{i=1}^{n}1_{\hat{e}_i\le z}, \text{and}\hat{g}^E_{h_1,a_1}(z)dz=d\hat{G}_{h_1,a_1}(z) \quad (11)$$

leads to the nonparametric location-scale estimator of $\int|z|g_{h_1, a_1}(z)dz$, $\int|z|\hat{g}^E_{h_1,a_1}(z)dz=n^{-1}\sum_{i=1}^{n}|\hat{m}(h_1,a_1)+\hat{\sigma}(h_1,a_1)\hat{e}_i|$. We show in Section 2·5 that the nonparametric location-scale estimator is consistent and asymptotically normal.

Data for estimating optimal sequential decision rules are typically expensive to collect, so samples are seldom large and parametric mean and variance function models are more useful than nonparametric models. Thus we assume that $m(h_1, a_1) = m(h_1, a_1; \theta)$ and $\sigma(h_1, a_1) = \sigma(h_1, a_1; \gamma)$ for some $\theta \in \mathbb{R}^{p_m}$ and $\gamma \in \mathbb{R}^{p_s}$. Similarly, we assume that $L(h_1, a_1) = L(h_1, a_1; \alpha)$, $\alpha \in \mathbb{R}^{p_L}$. For the results in Sections 3 we completed specification of the IQ-learning algorithm in Section 2·3 as follows. Steps IQ2a and IQ2b are amended to:

IQ2a Set $L\hat{}(h_1, a_1) = L(h_1, a_1; \alpha)\hat{}$, where

$$\hat{\alpha}=\arg\ \min_\alpha\sum_{i=1}^{n}\{\hat{\mu}(H_{2,i}) - L(H_{1,i}, A_{1,i};\alpha)\}^2$$

IQ2bi Set $m\hat{}(h_1, a_1) = m(h_1, a_1;\theta)\hat{}$, where

$$\hat{\theta}=\arg\ \min_\theta\sum_{i=1}^{n}\{\hat{\Delta}(H_{2,i}) - m(H_{1,i}, A_{1,i};\theta)\}^2$$

IQ2bii Set $\sigma\hat{}(h_1, a_1) = \sigma(h_1, a_1; \gamma)\hat{}$ where

$$\hat{\gamma}=\arg\ \min_\gamma\sum_{i=1}^{n}\left\{\log|\hat{\Delta}(H_{2,i}) - m(H_{1,i}, A_{1,i};\hat{\theta})| - \log\sigma(H_{1,i}, A_{1,i};\gamma)\right\}^2.$$

IQ2biii Set $\hat{g}_{h_1, a_1}$ to either $\hat{g}^N_{h_1,a_1}$ in (9) or to $\hat{g}^E_{h_1,a_1}$ in (11).

We have used a simple model for the conditional variance in step IQ2bii; for a discussion of other conditional variance estimators and their asymptotic properties see Carroll and Ruppert (1988).

2·5. Asymptotic Theory

Asymptotic distribution theory for $\hat{Q}_2^{\mathrm{IQ}}(h_1, a_1)$ is covered by standard results for linear regression, so we address only the asymptotic distribution of $\hat{Q}_1^{\mathrm{IQ}}(h_1, a_1)$ for the particular parametric estimators defined in the IQ-learning algorithm. Define the population residuals

$$\hat{a}\check{D}\check{G}(H_2, H_1, A_1;\theta, \gamma, \beta_2)=\{\Delta(H_2;\beta_2) - m(H_1, A_1;\theta)\}/\sigma(H_1, A_1;\gamma)$$

and the population parameters:

$$\begin{aligned}\beta_2^*&=\operatorname{argmin}_{\beta_2} E\{Y - Q_2(H_2, A_1;\beta_2)\}^2,\\ \theta^*&=\operatorname{argmin}_{\theta} E\{\Delta(H_2;\beta_2^*) - m(H_1, A_1;\theta)\}^2,\\ \gamma^*&=\operatorname{argmin}_{\gamma} E\{\log|\Delta(H_2;\beta_2^*) - m(H_1, A_1;\theta^*)| - \log\sigma(H_1, A_1;\gamma)\}^2,\\ \alpha^*&=\operatorname{argmin}_{\alpha} E\{\mu(H_2;\beta_2^*) - L(H_1, A_1;\alpha)\}^2.\end{aligned}$$

Let $\hat{\theta}, \hat{\gamma}, \hat{\beta}_2$, and $\hat{\alpha}$ denote $n^{1/2}$-consistent estimators of their population analogs θ^*, γ^*, β_2^*, and α^*. For $x \in \mathbb{R}^p$, let $\mathscr{B}_d(x)$ denote a ball of radius d centered at x, and let E_n denote the empirical expectation operator so that $E_n f = n^{-1}\sum_{i=1}^n f(H_{1,i}, A_{1,i}, H_{2,i}, A_{2,i}, Y_i)$. The asymptotic results are stated in terms of the seven centered statistics:

$$\Delta_L=L(h_1, a_1;\hat{\alpha})-L(h_1, a_1;\alpha^*), \Delta_m=m(h_1, a_1;\hat{\theta})-m(h_1, a_1;\theta^*), \Delta_\beta=\hat{\beta}_2-\beta_2^*,$$
$$\Delta_\sigma=\sigma(h_1, a_1;\hat{\gamma}) - \sigma(h_1, a_1;\gamma^*), \Delta_\theta=\hat{\theta} - \theta^*, \Delta_\gamma=\hat{\gamma} - \gamma^*,$$
$$\Delta_{\hat{a}\check{D}\check{G}}=E_n\{|m(h_1, a_1;\theta^*)+\sigma(h_1, a_1;\gamma^*)\hat{a}\check{D}\check{G}(H_2, H_1, A_1;\theta^*, \gamma^*, \beta_2^*)|\}-E\{|m(h_1, a_1;\theta^*)+\sigma(h_1, a_1;\gamma^*)\hat{a}\check{D}\check{G}(H_2, H_1, A_1;\theta$$

The following assumptions are used to establish the limit theory for IQ-learning:

(A1N) $n^{1/2}$ $(\Delta_L, \Delta_m, \Delta_\sigma)$ is asymptotically $N\{0, \Sigma_N(h_1, a_1)\}$;

(A1E) $n^{1/2}$ $(\Delta_L, \Delta_\theta, \Delta_\gamma, \Delta_\beta, \Delta_\varepsilon)$ is asymptotically $N\{0, \Sigma_E(h_1, a_1)\}$;

(A2) let k_1, k_2, and k_3 denote fixed positive constants, the class of functions

$$\mathscr{F}=\Big\{f(H_2, H_1, A_1;\theta, \gamma, \beta_2)=|m(h_1, a_1;\theta)+\sigma(h_1, a_1;\gamma)\hat{a}\check{D}\check{G}(H_2, H_1, A_1;\theta, \gamma, \beta_2)|:$$
$$\theta \in \mathscr{B}_{k_1}(\theta^*), \gamma \in \mathscr{B}_{k_2}(\gamma^*), \beta_2 \in \mathscr{B}_{k_3}(\beta_2^*)\Big\},$$

is a P-measurable Donsker class with a square-integrable envelope. In addition, $J(\theta, \gamma, \beta_2) = Ef(H_2, H_1, A_1; \theta, \gamma, \beta_2)$ is continuously differentiable with a bounded derivative in a neighborhood of $(\theta^*, \gamma^*, \beta_2^*)$;

(A3) the random variable $\mathscr{E}(H_2, H_1, A_1;\theta^*, \gamma^*, \beta_2^*)$ has a continuously differentiable density κ with derivative $\kappa'(z)$ satisfying $|\int z^2\kappa'(z)dz| < \infty$.

These assumptions are relatively mild with (A1N), (A1E), and the first part of (A2) verifiable using standard techniques, e.g., the multivariate central limit and Donsker preservation theorems (see Kosorok, 2008, and the Supplementary Material). Assumption (A3) is generally more difficult to verify, but its validity can be roughly assessed using the observed residuals $\hat{e}_1,\ldots,\hat{e}_n$. We have not sought the most general assumptions, but rather a set of simple assumptions that illustrate what is needed for the IQ-learning estimator to be well-behaved.

The first result states the asymptotic normality of the normal and nonparametric location-scale IQ-learning estimators of the first-stage Q-function. Let $\hat{Q}_1^{\mathrm{IQ},N}(h_1,a_1)$ and $\hat{Q}_1^{\mathrm{IQ},E}(h_1,a_1)$ denote the normal and nonparametric location-scale estimators, respectively, so that

$$\hat{Q}_1^{\mathrm{IQ},N}(h_1,a_1)=L(h_1,a_1;\hat{\alpha})+\int |z|\hat{g}^N_{h_1,a_1}(z)dz,\ \hat{Q}_1^{\mathrm{IQ},E}(h_1,a_1)=L(h_1,a_1;\hat{\alpha})+\int |z|\hat{g}^E_{h_1,a_1}(z)dz.$$

Define $I(\upsilon,t,s)=\upsilon+s^{-1}\int|z|\varphi\{(z-t)/s\}dz$. Let $I^*(h_1,a_1)$ denote $I\{L(h_1,a_1;\alpha^*), m(h_1,a_1;\theta^*), \sigma(h_1,a_1;\gamma^*)\}$. The following is proved in the Supplementary Material.

Theorem 1 (Asymptotic normality.). *Let h_1 and a_1 be fixed.*

1. *Assume (A1N). Then*

$$n^{1/2}\left[\hat{Q}_1^{\mathrm{IQ},N}(h_1,a_1)-L(h_1,a_1;\alpha^*)-\frac{1}{\sigma(h_1,a_1;\gamma^*)}\int|z|\phi\left\{\frac{z-m(h_1,a_1;\theta^*)}{\sigma(h_1,a_1;\gamma^*)}\right\}dz\right]$$

   converges in distribution to $N\{0,\nabla I^*(h_1,a_1)^\top\Sigma_N(h_1,a_1)\nabla I^*(h_1,a_1)\}$.

2. *Assume (A1E), (A2), and (A3). Then*

$$n^{1/2}\left[\hat{Q}_1^{\mathrm{IQ},E}(h_1,a_1)-L(h_1,a_1;\alpha^*)-\frac{1}{\sigma(h_1,a_1;\gamma^*)}\int|z|\kappa\left\{\frac{z-m(h_1,a_1;\theta^*)}{\sigma(h_1,a_1;\gamma^*)}\right\}dz\right]$$

   converges in distribution to

$$N[0,\{1,\nabla J(\theta^*,\gamma^*,\beta_2^*)^\top,1\}\textstyle\sum_E(h_1,a_1)\{1,\nabla J(\theta^*,\gamma^*,\beta_2^*)^\top,1\}^\top].$$

Theorem 1 shows that both location-scale estimators $\hat{Q}_1^N(h_1,a_1)$ and $\hat{Q}_1^E(h_1,a_1)$ are asymptotically normal under the stated conditions, which do not require correct specification of the IQ-learning models. Under (C1) and (C2) below, the IQ-learning models are correctly specified, and consistency and asymptotic normality of the IQ-learning estimators follow.

(C1) Let Z denote a standard normal random variable, then $\Delta(H_2;\beta_2^*)=m(H_1,A_1;\theta^*)+\sigma(H_1,A_1;\gamma^*)$.

(C2) Let W be a random variable with density $\kappa(\cdot)$, then $\Delta(H_2;\beta_2^*)=m(H_1,A_1;\theta^*)+\sigma(H_1,A_1;\gamma^*)W$.

The following results are direct consequences of Theorem 1 and we omit their proofs.

Corollary 1. *Assume* $Q_2(h_2,a_2)=Q_2(h_2,a_2;\beta_2^*)$ *and* $E\{\mu(H_2;\beta_2^*)|H_1=h_1,A_1=a_1\}=L(h_1,a_1;\alpha^*)$. *Then*

1. if (C1), $Q_1(h_1,a_1)=L(h_1,a_1;\alpha^*)+\frac{1}{\sigma(h_1,a_1;\gamma^*)}\int|z|\phi\left\{\frac{z-m(h_1,a_1;\theta^*)}{\sigma(h_1,a_1;\gamma^*)}\right\}dz$;
2. if (C2), $Q_1(h_1,a_1)=L(h_1,a_1;\alpha^*)+\frac{1}{\sigma(h_1,a_1;\gamma^*)}\int|z|\kappa\left\{\frac{z-m(h_1,a_1;\theta^*)}{\sigma(h_1,a_1;\gamma^*)}\right\}dz$.

Theorem 2. *Assume* (*A1N*) *and the conditions of Theorem* 1. *Let* h_1 *and* a_1 *be fixed.*

1. *If (C1) then* $n^{1/2}\{\hat{Q}_1^{\mathrm{IQ},N}(h_1,a_1)-Q_1(h_1,a_1)\}$ *converges in distribution to* $N\{0, \nabla I^*(h_1,a_1)^\top\Sigma_N(h_1,a_1)\nabla I^*(h_1,a_1)\}$.
2. *If (A1E), (A2), (A3) and (C2) then* $n^{1/2}\{\hat{Q}_1^{\mathrm{IQ},E}(h_1,a_1)-Q_1(h_1,a_1)\}$ *converges in distribution to* $N[0,\{1,\nabla J(\theta^*,\gamma^*,\beta_2^*)^\top,1\}\Sigma_E(h_1,a_1)\{1,\nabla J(\theta^*,\gamma^*,\beta_2^*)^\top,1\}^\top]$.

Remark 1. Theorem 2 can be used to construct asymptotically valid confidence intervals for the first-stage Q-function for fixed patient history h_1 and first-stage treatment a_1. This is a notoriously difficult task with Q-learning (Laber et al., 2014). In practice, due to the complexity of the variance terms, the bootstrap may be preferred. Remarks about how to extend the above results to obtain bootstrap consistency are made in the Supplementary Material.

Remark 2. As noted in the introduction, IQ-learning does not alleviate the inherent nonregularity present in sequential decision making problems; see Robins (2004), Laber et al. (2014), and Chakraborty et al. (2010). However, IQ-learning is consistent for a nonregular scenario of interest, the so-called global null in which there is no treatment effect for any patients at the second stage, i.e., $\Delta(H_2;\beta_2^*)=0$ almost surely. To see this, note that assuming (A1N), (C1) holds with $m(H_1, A_1; \theta^*) = 0$ with $\sigma(H_1, A_1; \gamma^*) \to 0$ almost surely. Part 1 of Corollary 1 depends only on part 1 of Theorem 1 and part 1 of Theorem 2. For the more complex case in which $0<\mathrm{pr}\{\Delta(H_2;\beta_2^*)=0\}<1$ we conjecture that using a mixture of normals to estimate $g_{h_1,a_1}(\cdot)$ may lead to improved small–sample performance.

3. Monte Carlo Results

We compare the small-sample performance of IQ- and Q-learning in using average value of the learned treatment regimes, integrated mean squared error of the first-stage Q-function, and coverage and width of 95% nonparametric bootstrap confidence intervals for the first-stage Q-function. A key advantage of IQ-learning is its compatibility with common model building steps, which we illustrate in the Supplementary Material with a study of the power to detect nonlinear effects in the first-stage Q-function using IQ-learning and Q-learning. Software implementing the IQ-learning estimators is available as part of the iqLearn package on the comprehensive R network (cran.us.r-project.org/). The Monte Carlo results show that for the class of generative models we consider, IQ-learning generally performs

better than Q-learning in terms of value, integrated mean squared error, coverage, and width of the confidence intervals. Simulations in the Supplementary Material show that IQ-learning also has higher power to detect nonlinear effects.

Simulations in this section use data from the following class of generative models:

$$X_1 \sim \text{Normal}_p\{0.1, \Omega_{AR_1}(0.5)\}, A_t \sim \text{Uniform}\{-1, 1\}, t=1,2,\ X_2=(1.5-0.5A_1)X_1+\zeta_{A_1}\xi, Y=H_{2,0}^\top\beta_{2,0}+A_2H_{2,1}^\top\beta_{2,1}+\phi,$$

where $\{\Omega_{\text{AR}_1}(0.5)\}_{i,j} = (0.5)^{|i-j|}$, $H_{2,0}=H_{2,1}=(1, X_2^\top, A_1, A_1X_2^\top)^\top$, and $\zeta_{\text{A}_1} = (1.5 + 0.5A_1)^{1/2}$. Thus, the class is indexed by the dimension p, the distributions of ξ and φ, and the coefficient vectors $\beta_{2,0}$ and $\beta_{2,1}$. Here we fix $p = 4$; results for $p = 8$ are similar and are provided in the Supplementary Material. We consider $\xi \sim \text{Normal}_p(0, I_p)$. We fix the main effect parameter $\beta_{2,0}$ and vary the second-stage treatment effect size by scaling $\beta_{2,1}$ as follows: $\beta_{2,0} = 1_{2p+2}/\|1_{2p+2}\|$, $\beta_{2,1}=C\{-0.25(1_{p+1}^\top), 1_{p+1}^\top\}^\top/\|\{-0.25(1_{p+1}^\top), 1_{p+1}^\top\}\|$, where C ranges over a grid from 0 to 2, and 1_d denotes a d-dimensional vector of 1s. In addition, we fix the theoretical R^2 of the second-stage regression model at 0.6 by specifying $\phi \sim \text{Normal}\{0, \sigma_\phi^2(C)\}$ and solving for the variance $\sigma_\phi^2(C)$ that yields the desired R^2. Additional simulations, provided in the Supplementary Material, show results for $R^2 = 0.4$, 0.8 and non-normal error distributions for ξ;

We consider linear working models for the mean and log variance functions

$$\begin{array}{llll} Q_2(h_2, a_2;\beta_2) & =h_2^\top\beta_{2,0}+a_2h_2^\top\beta_{2,1}, & Q_1(h_1, a_1;\beta_1) & =h_1^\top\beta_{1,0}+a_1h_1^\top\beta_{1,1}, \\ L(h_1, a_1;\alpha) & =h_1^\top\alpha_0+a_1h_1^\top\alpha_1, & m(h_1, a_1;\theta) & =h_1^\top\theta_0+a_1h_1^\top\theta_1, \\ \log\{\sigma(h_1, a_1;\gamma)\} & =h_1^\top\gamma_0+a_1h_1^\top\gamma_1, & & \end{array}$$

where now $H_1=(1, X_1^\top)^\top$. In addition to Q-learning with linear working models, we include results using support vector regression using a Gaussian kernel (Zhao et al., 2011) to estimate both Q-functions. We compare the two versions of Q-learning with two IQ-learning estimators that differ in the estimation of $g_{h_1,a_1}(\cdot)$ and the model for $\sigma(h_1, a_1; \gamma)$: a normal estimator $\hat{g}^N_{h_1,a_1}(\cdot)$ of the residual distribution and a restricted variance model, $\log\{\sigma(h_1, a_1; \gamma)\} = \gamma_0 + a_1\gamma_1$, that depends only on treatment; and a nonparametric estimator $\hat{g}^E_{h_1,a_1}(\cdot)$ of the residual distribution with a log-linear variance model that depends on h_1 and a_1. When $\xi \sim \text{Normal}_p(0, I_p)$, both these estimators are correctly specified. Q-learning is always correctly specified at the second stage but only correctly specified at the first stage when $C = 0$ and hence $\beta_{2,1} = 0$.

Results are based on a training set of size $n = 250$ and $M = 2{,}000$ Monte Carlo data sets for each generative model. Additional results for $n = 500$ are provided in the Supplementary Material. For the nonparametric IQ-learning estimator, which is always correctly specified, the true Q-functions and subsequent optimal regime are estimated using a test set of 10,000 observations. Recall that the value, $E^\pi(Y)$, of an arbitrary policy π is the expected outcome if all patients are assigned treatment according to π, that is, $E^\pi(Y) = E(E[E\{Y \mid H_2, a_2\} \mid$

$H_1, a_1]$) evaluated at $a_2 = \pi_2(H_2)$ and $a_1 = \pi_1(H_1)$. For a given training set of size n and an algorithm that produces an estimated optimal policy, say $\hat{\pi}$, we define the average value as $E\{E^{\hat{\pi}}(Y)\}$, where the outer expectation is taken over all training sets of size n. We estimate the average value of the IQ- and Q-learning estimators using a test set of size 10,000 to estimate the inner expectation and 2,000 Monte Carlo replications to estimate the outer expectation. We compare average values of the learned IQ and Q regimes to the value of the true optimal regime and present the proportion of optimal value obtained. For an estimator $\hat{Q}_1(h_1, a_1)$ of the first-stage Q-function, $Q_1(h_1, a_1)$, define the integrated mean squared error as $E[\{\hat{Q}_1(H_1, A_1) - Q_1(H_1, A_1)\}^2]$, where the expectation is taken over the joint distribution of (H_1, A_1) as well as the training data.

Confidence intervals for $Q_1(h_1, a_1)$ based on IQ-learning and Q-learning estimators are formed by bootstrapping the respective estimators and taking percentiles. For example, if $\hat{l}$ and $\hat{u}$ denote the $100 \times \eta/2$ and $100 \times (1 - \eta/2)$ percentiles of the bootstrap distribution of $Q^{I\hat{Q}}(h_1, a_1)$ based on 1,000 bootstrap resamples, then the $100 \times (1 - \eta)\%$ confidence interval is given by $(2Q^{I\hat{Q}}(h_1, a_1) - \hat{u}, 2Q^{I\hat{Q}}(h_1, a_1) - \hat{l})$. Bootstrap intervals of this form are sometimes referred to as hybrid bootstrap confidence intervals (Efron and Tibshirani, 1993). Coverage and width of the foregoing confidence intervals are estimated using 2,000 Monte Carlo replications with a new instance (h_1, a_1) of (H_1, A_1) drawn for each replication. Figure 2 displays the results from this simulation, where $\xi \sim \text{Normal}_p(0, I_p)$.

Figure 2 indicates that the IQ-learning estimators perform better than both Q-learning estimators. Although some gains are achieved using the more flexible support vector regression version of Q-learning, the far left panel indicates that IQ-learning attains higher average value than the Q-learning algorithms across most values of C. The second panel from the left shows that the IQ-learning reduces integrated mean squared error the most, with greater reduction as the second-stage effects increase. IQ-learning also demonstrates a large improvement over linear Q-learning in terms of the coverage of 95% confidence intervals for $\hat{Q}_1(h_1, a_1)$, as seen in the third panel. The poor coverage of linear Q-learning is attributed to bias, whereas IQ-learning is consistent because the regularity conditions given in Section 2.4 for Theorem 2 hold for the generative models in this section. Thus, IQ-learning estimators come close to achieving the nominal level. The average widths of the confidence intervals are similar for linear Q-learning and IQ-learning, as illustrated by the far right panel. Support vector regression Q-learning greatly improves coverage compared to linear Q-learning, but at the expense of wider intervals. Results from the simulation where the elements of ξ are generated independently from a t_5 distribution are included in the Supplementary Material and appear similar to those in Fig. 2, suggesting the normal IQ-learning estimator is robust to slight misspecification of the residual distribution.

Next, we consider the model

$$\begin{array}{lll} X_1 \sim \text{Normal}(-2, 1), & \xi \sim \text{Normal}(0, 1), & A_t \sim \text{Uniform}\{-1, 1\}, t = 1, 2, \\ X_2 = X_1 + \xi, & \phi \sim \text{Normal}(0, 1), & Y = H_{2,0}^{\top}\beta_{2,0} + A_2 H_{2,1}^{\top}\beta_{2,1} + \phi, \end{array}$$

where $H_{2,0} = H_{2,1} = (1, X_1, A_1, X_1A_1, X_2)^\top$, $\beta_{2,0} = (3, -1, 0.1, -0.1, -0.1)^\top$, and $\beta_{2,1} = C(-6, -2, 5, 3, -0.2)^\top$. We use this example to illustrate a scenario where IQ-learning achieves a large gain in value over Q-learning with linear models. Since the predictor X_1 is univariate, we can visualize which patients are treated differently by IQ-learning compared to Q-learning. The left plot in Figure 3 was obtained by deriving the true first stage Q-function, for which IQ-learning is consistent in this case, and comparing the true first-stage rule to the rule recommended by Q-learning with linear models. For each combination of (X_1, C), the plot shows whether or not Q-learning makes the correct treatment decision. With this generative model, Q-learning assigns the wrong treatment to approximately half the population for a wide range of effect sizes because the true first-stage Q-function is nonlinear in X_1. In contrast, with a sufficiently large sample size, IQ-learning treats all patients according to the optimal rule. A remark explaining the observed pattern is included in Section 6 of the Supplementary Material. Consequently, IQ-learning achieves higher average value than Q-learning, as displayed in the right plot of Fig. 3. Results are based on $n = 250$ training set samples and $M = 1{,}000$ Monte Carlo data sets. In this scenario, both IQ-learning and support vector regression Q-learning reach gains in optimal value attained of approximately 15% as the second-stage effect size grows.

4. Application to STAR*D

Sequenced Treatment Alternatives to Relieve Depression (STAR*D; Fava et al., 2003; Rush et al., 2004) is a sequentially randomized study of major depressive disorder. A key feature of this trial is that patients experiencing early remission of symptoms were exempt from future randomization, complicating the analysis. We use a subset of the STAR*D data to illustrate how IQ-learning can be used to estimate an optimal dynamic treatment regime in the presence of responder-status dependent designs. There were four stages in the trial, but each patient received Citalopram in the first stage, and thus, there was no randomization. Since our aim is to demonstrate how to learn a regime using the IQ-learning machinery, we opt to perform a complete-case analysis and consider only the first two of three randomized stages. We refer to the second and third stages as stages one and two, respectively. At each stage, treatments can be categorized as a Selective Serotonin Reuptake Inhibitor or not; this is the binary treatment variable in our analysis. The first-line treatment Citalopram given in the non-randomized stage is in the class of Selective Serotonin Reuptake Inhibitors.

We use a measure of efficacy, the Quick Inventory of Depression Symptomatology score, as the outcome (Rush et al., 2004); side-effects or other competing outcomes could be accommodated using set-valued treatment regimes (Lizotte et al., 2012; Laber et al., 2013) or composite outcomes (Wang et al., 2012). The depression score ranges from 0 to 27 with higher values corresponding to more severe negative symptoms. To be consistent with our development, we recode these scores by subtracting them from 27; thus, higher values correspond to better clinical outcomes. The depression score was recorded at multiple time points throughout each stage, intermediate depression scores used as predictors in our analysis are also recoded (Rush et al., 2004, Schulte et al., 2012). At each stage, patients experiencing remission left the study. Remission was defined as a depression score of 5 or less, or greater than 21 after recoding. Henceforth, we refer to patients who achieved remission and left after stage one as responders and second-stage participants as non-

responders. The data we use here consist of $n = 795$ patient trajectories with complete information, not including 481 patients who dropped out for reasons other than remission. The variables composing each trajectory are given in Table 1 of the Supplementary Material.

Of the 795 total patients, 329 were non-responders in the first stage and continued on to the second stage. We define our primary outcome as $Y = RY_1 + (1 - R)(Y_1 + Y_2)/2$. That is, Y is taken to be Y_1 for patients who left the study after stage one, and Y is taken to be the average of the depression scores measured at the end of the first and second stages for non-responders. Our responder-status version of IQ-learning is based on the Q-learning implementation described in Schulte et al. (2012). The first-stage history vector contains all information available prior to the first-stage treatment randomization. Thus, $H_1 = (X_{1,1}, X_{1,2})^\top$. The second-stage history is $H_2 = (X_{1,1}, X_{1,2}, A_1, Y_1, R, X_{2,1}, X_{2,2})^\top$, which contains all information observed before the second-stage treatment assignment. The second-stage Q-function is $Q_2(H_2, A_2) = E(Y \mid H_2 = h_2, A_2 = a_2) = RY_1 + (1 - R)\{Y_1 + E(Y_2 \mid H_2 = h_2, A_2 = a_2)\}/2$, where we have used the fact that R and Y_1 are contained in H_2. Thus the first step in the IQ-learning algorithm, and in Q-learning, is to specify and fit a model for $E(Y_2 \mid H_2 = h_2, A_2 = a_2)$. Defining the second-stage summary vectors as $H_{2,0} = H_{2,1} = (1, X_{2,1}, X_{2,2})^\top$, we consider a working model of the form $E(Y_2|H_2{=}h_2, A_2{=}a_2){=}h_{2,0}^\top\beta_{2,0}{+}a_2h_{2,1}^\top\beta_{2,1}$, that we fit via least squares. Standard regression diagnostics based on the residuals to not indicate any major departures from the usual linear modeling assumptions.

In our subset of data, all non-responders who did not receive a Selective Serotonin Reuptake Inhibitor in stage one did not receive one in stage two. Table 2 of the Supplementary Material provides the number of patients assigned to each treatment strategy. We define $\tilde{Y} = \arg\max_{a_2} Q_2(H_2, a_2)R + \{Q_2(H_2, -1)(1 - A_1) + \arg\max_{a_2} Q_2(H_2, a_2)(1 + A_1)\}(1 - R)/2$. Thus, $\tilde{Y}$ is $Q_2(H_2, -1)$ for non-responders who received $A_1 = -1$ and $\arg\max_{a_2} Q_2(H_2, a_2)$ otherwise. With this working model, after substituting in $Q_2(H_2, a_2)$ and simplifying,

$$\begin{aligned} Q_1(H_1, A_1) &= E(RY_1|H_1, A_1) \\ &+ \frac{1}{2}E\{(1 - R)Y_1|H_1, A_1\} + \frac{1}{2}E\{(1 - R)H_{2,0}^\top\beta_{2,0}|H_1, A_1 - \frac{1 - A_1}{4}E\{(1 - R)H_{2,1}^\top\beta_{2,1}|H_1, A_1\} \\ &+ \frac{1+A_1}{4}E\{(1 - R)|H_{2,1}^\top\beta_{2,1}||H_1, A_1\}. \end{aligned} \tag{12}$$

The Q-learning algorithm models $E(\tilde{Y} \mid H_1, A_1)$ directly. The left panel in Figure 4 shows a scatterplot of the pseudo-response $\tilde{Y}$ against baseline depression score by first-stage

treatment. Cubic smoothing spline fits to the data are indicated by solid gray and dashed black lines for $A_1 = 1$ and $A_1 = -1$, respectively. These fits appear approximately linear; however, the variance appears non-constant across baseline depression score, and there is clear separation between the responder and non-responder groups.

We can write the first three expectation terms in (12) as:

$$\begin{aligned} E(RY_1|H_1, A_1) &= E(Y_1|H_1, A_1, R{=}1)\mathrm{pr}(R{=}1|H_1, A_1), \\ E\{(1-R)Y_1|H_1, A_1\} &= E(Y_1|H_1, A_1, R{=}0)\mathrm{pr}(R{=}0|H_1, A_1), \\ E\left\{(1-R)H_{2,0}^\top\beta_{2,0}|H_1, A_1\right\} &= E(H_{2,0}^\top\beta_{2,0}|H_1, A_1, R{=}0)\mathrm{pr}(R{=}0|H_1, A_1). \end{aligned}$$

We estimate the right-hand conditional expectations by fitting three separate linear regressions, and we use logistic regression to estimate pr($R = r \mid H_1, A_1$, $r = 0, 1$. In particular, we specify linear models of the form

$E(Y_1|H_1{=}h_1, A_1{=}a_1, R{=}r){=}h_{1,0}^\top\lambda_{r,0}{+}a_1h_{1,1}^\top\lambda_{r,1}$ for $E(Y_1 \mid H_1, A_1, R = r)$, where $H_{1,0} = H_{1,1} = (1, X_{1,1}, X_{1,2})^\top$. We posit the model

$E\left(H_{2,0}^\top\beta_{2,0}|H_1{=}h_1, A_1{=}a_1, R{=}0\right){=}h_{1,0}^\top\alpha_0{+}a_1h_{1,1}^\top\alpha_1$ for the main-effect term, which we fit with least squares using only the non-responder data. The middle and right plots in Fig. 4 display scatterplots of the non-responder realizations of the main-effect term and contrast function, respectively, against baseline depression score by first-stage treatment. Cubic smoothing spline fits to the data appear mostly linear. For the logistic regression, we fit the model $\mathrm{logit}\{\mathrm{pr}(R{=}1|H_1{=}h_1, A_1{=}a_1)\}{=}h_{1,0}^\top\delta_0{+}a_1h_{1,1}^\top\delta_1$.

Finally, we must obtain estimates of $E\{(1-R)|H_{2,1}^\top\beta_{2,1}||H_1, A_1\}$ and $E\{(1-R)H_{2,1}^\top\beta_{2,1}|H_1, A_1\}$ in equation (12). Notice that

$$E\left\{(1-R)|H_{2,1}^\top\beta_{2,1}||H_1, A_1\right\} = \mathrm{pr}(R{=}0|H_1, A_1)\int |z|g_{H_1,A_1,R=0}(z)dz.$$

We can use the IQ-learning machinery to obtain an estimate of $g_{H_1, A_1, R=0}(\cdot)$. The logistic regression previously described provides an estimate of pr($R = 0 \mid H_1, A_1$). To estimate $g_{H_1, A_1, R=0}(\cdot)$, we first specify a model for the mean of the contrast function for non-responders. We posit the model

$$E\left\{\hat{\Delta}(H_2;\hat{\beta}_2)|H_1, A_1\right\} = H_{1,0}^\top\beta_0 + A_1H_{1,1}^\top\beta_1, \quad (13)$$

where $H_{1,0} = (1, X_{1,1}, X_{1,2})^\top$ and $H_{1,1} = (1, X_{1,1}, X_{1,2})^\top$. This model is also used along with the logistic regression model to estimate

$E\{(1-R)H_{2,1}^\top\beta_{2,1}|H_1, A_1\}{=}E(H_{2,1}^\top\beta_{2,1}|H_1, A_1, R{=}0) \times \mathrm{pr}(R{=}0|H_1, A_1)$, the fourth expectation term in (12). We fit model (13) using least squares. Exploratory analysis suggests that a constant variance assumption is reasonable, so we standardize the residuals of the mean fit using the sample standard deviation. A normal quantile-quantile plot of the standardized residuals suggests heavier tails than would be expected from a normal

distribution. Thus, we opt to use the nonparametric density estimator described in Section 2.3.

Assembling the foregoing estimates of the four terms in equation (12) yields

$$\begin{aligned}\hat{Q}_1^{IQ}(h_1, a_1;\hat{\theta}_1) &= (h_{1,0}^\top \hat{\lambda}_{1,0} \\ &+ a_1 h_{1,1}^\top \hat{\lambda}_{1,1}) \mathrm{expit}(h_{1,0}^\top \hat{\delta}_0 \\ &+ a_1 h_{1,1}^\top \hat{\delta}_1) \\ &+ \Big\{ 1 - \mathrm{expit}(h_{1,0}^\top \hat{\delta}_0 \\ &+ a_1 h_{1,1}^\top \hat{\delta}_1) \Big\{ \frac{1}{2}(h_{1,0}^\top \hat{\lambda}_{0,0} \\ &+ a_1 h_{1,1}^\top \hat{\lambda}_{0,1}) \\ &+ \frac{1}{2}(h_{1,0}^\top \hat{\alpha}_0 \\ &+ a_1 h_{1,1}^\top \hat{\alpha}_1) \\ &- \frac{1}{4}(h_{1,0}^\top \hat{\beta}_0 \\ &+ a_1 h_{1,1}^\top \hat{\beta}_1) \\ &+ \frac{1}{4} \int |z| \hat{g}_{h_1, a_1, R=0}(z) dz \Big\},\end{aligned}$$

where $\hat{g}_{h_1, a_1, R=0}(\cdot)$ is the nonparametric density estimator described in Section 2.3 and $\hat{\theta}_1 = (\hat{\lambda}_{0,0}^\top, \hat{\lambda}_{0,1}^\top, \hat{\lambda}_{1,0}^\top, \hat{\lambda}_{1,1}^\top, \hat{\delta}_0^\top, \hat{\delta}_1^\top, \hat{\alpha}_0^\top, \hat{\alpha}_1^\top, \hat{\beta}_0^\top, \hat{\beta}_1^\top)^\top$.

The first-stage rule estimated by Q-learning treats all training data patients with a Selective Serotonin Reuptake Inhibitor. Roughly twelve percent of the these patients are not recommended a Selective Serotonin Reuptake Inhibitor by the estimated first-stage IQ-learning rule. Broadly summarizing the IQ-learning rule, patients with very low recoded depression scores after the non-randomized stage should switch from Citalopram, a Selective Serotonin Reuptake Inhibitor, to another drug not in the class of Selective Serotonin Reuptake Inhibitors. This rule recommends a different strategy for patients who respond poorly to the initial Selective Serotonin Reuptake Inhibitor; otherwise Selective Serotonin Reuptake Inhibitor treatment strategies should continue.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge grant support from the National Institutes of Health (Laber, Stefanski) and National Science Foundation (Stefanski). The authors thank the National Institutes of Health for providing the STAR*D data and Bibhas Chakraborty for his consultation on the analysis presented in Section 4.

References

Bellman, R. Dynamic Programming. Princeton: Princeton University Press; 1957.

Blatt D, Murphy SA, Zhu J. A-Learning for Approximate Planning. Ann Arbor, 1001. 2004:48109–42122.

Carroll, RJ.; Ruppert, D. Transformation and Weighting in Regression. New York: Chapman and Hall; 1988.

Chakraborty, B.; Moodie, EE. Statistical Methods for Dynamic Treatment Regimes. Springer; 2013. Statistical reinforcement learning; p. 31-52.

Chakraborty B, Murphy SA, Strecher VJ. Inference for Non-Regular Parameters in Optimal Dynamic Treatment Regimes. Statistical Methods in Medical Research. 2010; 19(3):317–343. [PubMed: 19608604]

Cook, RD.; Weisberg, S. Residuals and Influence in Regression. New York: Chapman and Hall; 1982.

Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. New York: Chapman and Hall; 1993.

Fava M, Rush AJ, Trivedi MH, Nierenberg AA, Thase ME, Sackeim HA, Quitkin FM, Wisniewski SR, Lavori PW, Rosenbaum JF, Kupfer DJ. STAR*D Investigators Group. Background and Rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Study. Psychiatric Clinics of North America. 2003; 26(1):457+. [PubMed: 12778843]

Henderson HV, Velleman PF. Building Multiple Regression Models Interactively. Biometrics. 1981; 37(2):391–411.

Henderson R, Ansell P, Alshibani D. Regret-regression for optimal dynamic treatment regimes. Biometrics. 2010; 66(4):1192–1201. [PubMed: 20002404]

Kosorok, MR. Introduction to Empirical Processes and Semiparametric Inference. New York: Springer; 2008.

Laber E, Lizotte D, Ferguson B. Set-valued dynamic treatment regimes for competing outcomes. Biometrics. 2013

Laber EB, Lizotte DJ, Qian M, Pelham WE, Murphy SA. Statistical Inference in Dynamic Treatment Regimes. Under Review. 2014

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

Lavori PW, Dawson R. Dynamic Treatment Regimes: Practical Design Considerations. Clinical Trials. 2004; 1(1):9–20. [PubMed: 16281458]

Lizotte DJ, Bowling M, Murphy SA. Linear fitted-q iteration with multiple reward functions. Journal of Machine Learning Research. 2012; 13:3253–3295. [PubMed: 23741197]

Moodie EE, Dean N, Sun YR. Q-learning: Flexible learning about useful utilities. Statistics in Biosciences. 2013:1–21.

Moodie EE, Richardson TS. Estimating optimal dynamic regimes: Correcting bias under the null. Scandinavian Journal of Statistics. 2010; 37(1):126–146.

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

Murphy SA. A Generalization Error for Q-Learning. Journal of Machine Learning Research. 2005a; 6(7):1073–1097. [PubMed: 16763665]

Murphy SA. An Experimental Design for the Development of Adaptive Treatment Strategies. Statistics in Medicine. 2005b; 24(10):1455–1481. [PubMed: 15586395]

Rich B, Moodie EE, Stephens DA, Platt RW. Model checking with residuals for g-estimation of optimal dynamic treatment regimes. The international journal of biostatistics. 2010; 6(2)

Robins, JM. Proceedings of the Second Seattle Symposium in Biostatistics. New York: Springer; 2004. Optimal Structural Nested Models for Optimal Sequential Decisions; p. 189-326.

Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA, Thase ME, Nierenberg AA, Quitkin FM, Kashner T, Kupfer DJ, Rosenbaum JF, Alpert J, Stewart JW, McGrath PJ, Biggs MM, Shores-Wilson K, Lebowitz BD, Ritz L, Niederehe G. STAR*D Investigators Group. Sequenced Treatment Alternatives to Relieve Depression (STAR*D): Rationale and Design. Controlled Clinical Trials. 2004; 25(1):119–142. [PubMed: 15061154]

Schulte PJ, Tsiatis AA, Laber EB, Davidian M. Q-and a-learning methods for estimating optimal dynamic treatment regimes. arXiv preprint arXiv:1202.4177. 2012

Sutton, RS.; Barto, AG. Reinforcement Learning: An Introduction. 1st edition. Cambridge: MIT Press, Cambridge, MA, USA; 1998.

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. Journal of the American Statistical Association. 2012; 107(498):493–508. [PubMed: 22956855]

Zhao Y, Zeng D, Socinski MA, Kosorok MR. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. Biometrics. 2011; 67(4):1422–1433. [PubMed: 21385164]

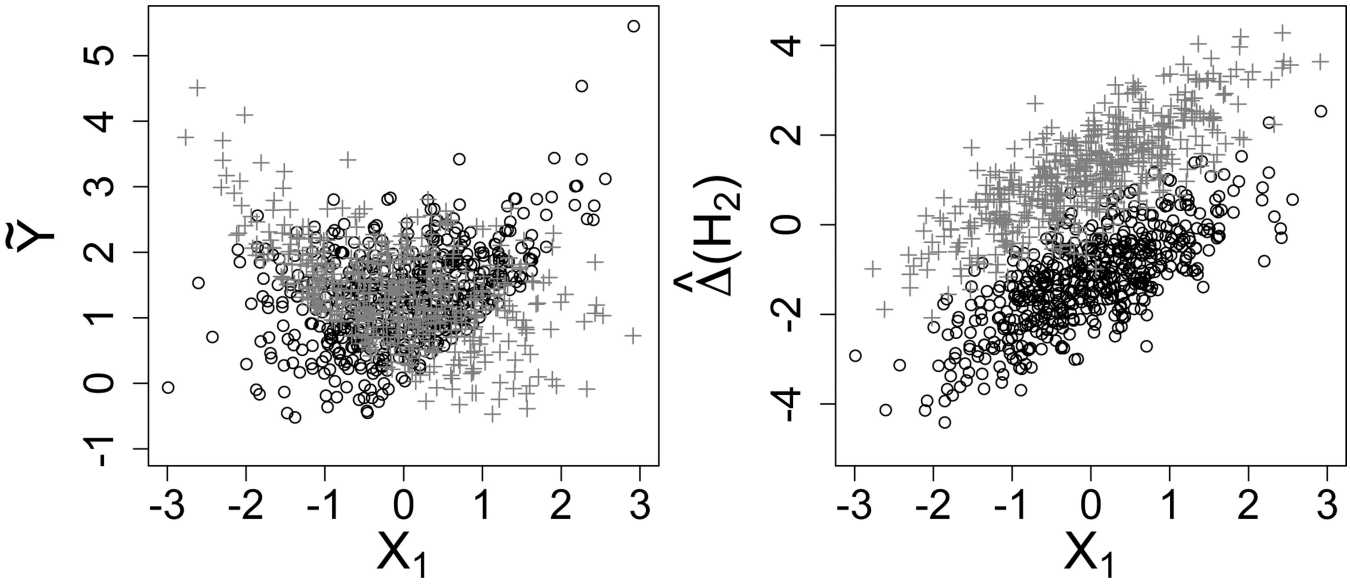

Fig. 1.

Scatterplots of $\tilde{Y}$ (left) and $\hat{\Delta}(H_2)$ (right) against X_1 for $A_1 = -1$ (black circles) and $A_1 = 1$ (grey crosses) for 1, 000 random samples from the toy model. Step 2 of the Q-learning algorithm requires modeling the data in the left plot; note the nonlinearity and heteroscedasticity. Data in the right plot must be modeled for IQ-learning; note the common analysis of covariance structure.

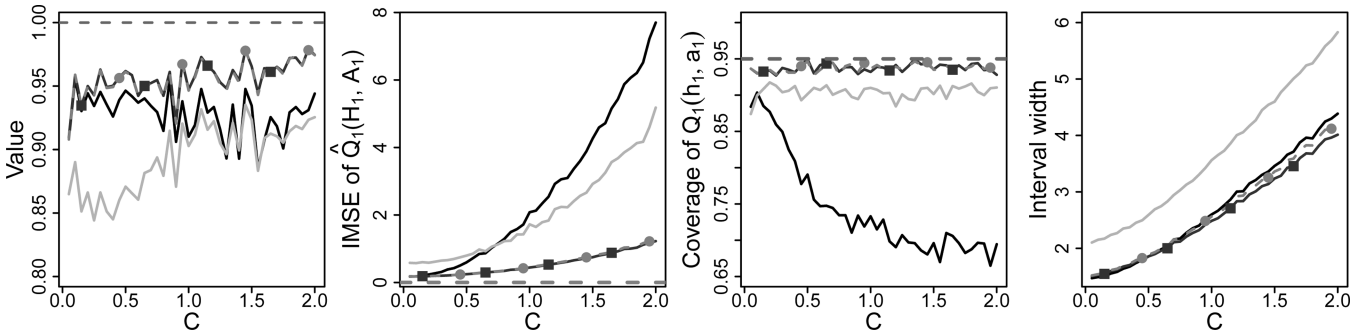

Fig. 2.

Performance of the normal IQ-learning estimator, nonparametric IQ-learning estimator, support vector regression Q-learning, and linear Q-learning given by gray lines with squares, gray lines with circles, light gray lines, and black lines, respectively. Left to Right: Average proportion of optimal value attained; integrated mean squared error of Q_1 estimates; coverage of 95% confidence intervals for Q_1; width of 95% confidence intervals for Q_1.

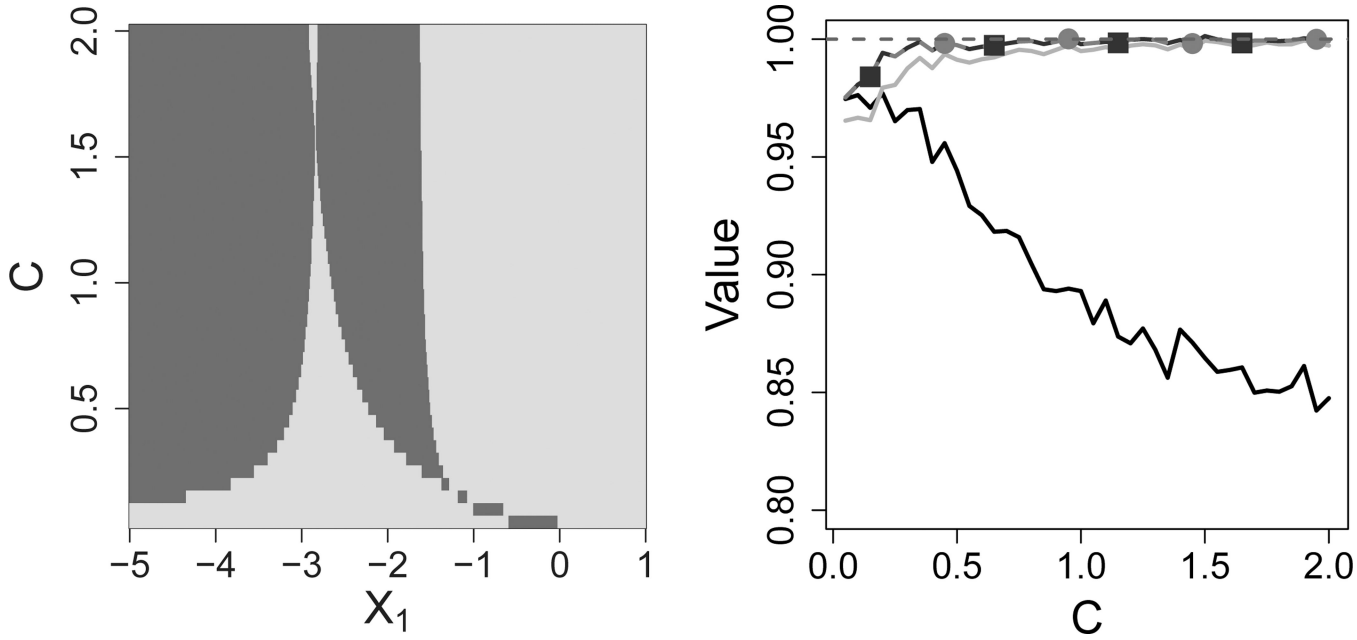

Fig. 3.

A scenario where IQ-learning achieves a large gain in value over Q-learning. The constant C determines the second-stage treatment effect size, from no treatment effects ($C = 0$) to large effects ($C = 2$). In the left panel, (X_1, C) pairs where linear Q-learning agrees and disagrees with the true first-stage rule are shown in dark and light gray, respectively, where X_1 is a normally distributed first-stage covariate. On the right, average proportion of optimal value attained by the normal IQ-learning estimator, nonparametric IQ-learning estimator, support vector regression Q-learning, and linear Q-learning regimes shown by gray lines with squares, gray lines with circles, light gray lines, and black lines, respectively.

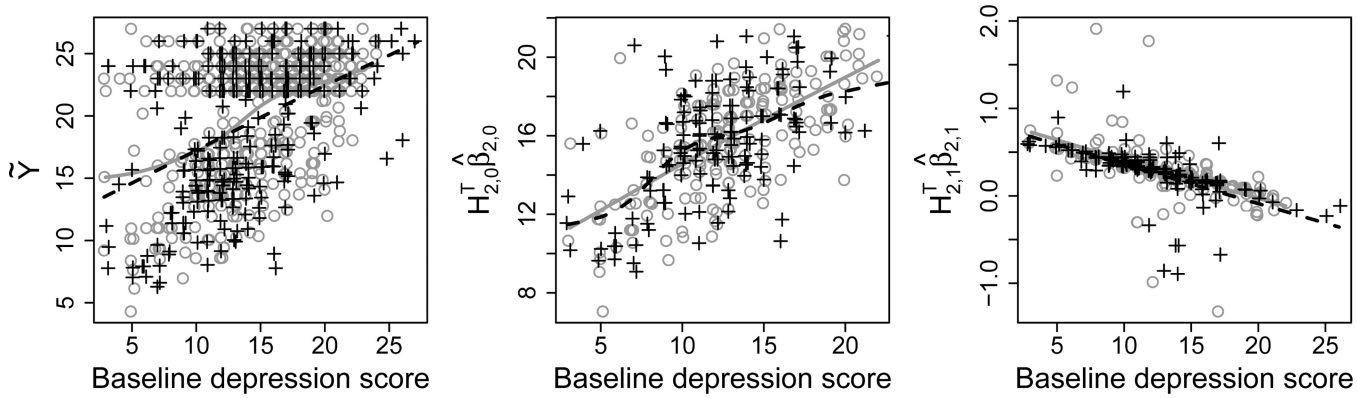

Fig. 4.

First-stage summaries of the STAR*D data. Gray circles and black crosses represent treatments $A_1 = 1$ and $A_1 = -1$, respectively. The left panel shows a scatterplot of $\tilde{Y}$ against baseline depression score by first-stage treatment. The middle panel contains a scatterplot of $H_{2,0}^{\top}\hat{\beta}_{2,0}$ against baseline depression score by first-stage treatment for non-responders. The right panel contains a scatterplot of $H_{2,1}^{\top}\hat{\beta}_{2,1}$ against baseline depression score by first-stage treatment for non-responders. Cubic smoothing spline fits to the data for $A_1 = 1$ and $A_1 = -1$ are represented by solid gray and dashed black lines, respectively.