Published in final edited form as:

Stat Med. 2016 June 15; 35(13): 2221–2234. doi:10.1002/sim.6859.

Q-learning Residual Analysis: Application to The Effectiveness of Sequences of Antipsychotic Medications for Patients with Schizophrenia

Ashkan Ertefaie^{1,*†}, Susan Shortreed², and Bibhas Chakraborty³

¹Department of Statistics, University of Pennsylvania, Philadelphia, PA, USA; Center for Pharmacoepidemiology Research and Training, University of Pennsylvania, Philadelphia, PA, USA

²Biostatistics Unit, GroupHealth Research Institute, Seattle, WA, USA

³Centre for Quantitative Medicine, Duke-NUS Graduate Medical School, Singapore; Department of Biostatistics, Columbia University, New York, NY, USA

Abstract

Q-learning is a regression-based approach that uses longitudinal data to construct dynamic treatment regimes, which are sequences of decision rules that use patient information to inform future treatment decisions. An optimal dynamic treatment regime is composed of a sequence of decision rules that indicate how to optimally individualize treatment using the patients' baseline and time-varying characteristics to optimize the final outcome. Constructing optimal dynamic regimes using Q-learning depends heavily on the assumption that regression models at each decision point are correctly specified; yet model checking in the context of Q-learning has been largely overlooked in the current literature. In this article, we show that residual plots obtained from standard Q-learning models may fail to adequately check the quality of the model fit. We present a modified Q-learning procedure that accommodates residual analyses using standard tools. We present simulation studies showing the advantage of the proposed modification over standard Q-learning. We illustrate this new Q-learning approach using data collected from a sequential multiple assignment randomized trial of patients with schizophrenia.

Keywords

Dynamic treatment regimes; Q-learning; Residual analysis; SMART designs

*Correspondence to: Department of Statistics, University of Pennsylvania, Philadelphia, PA, USA.

†ertefaie@wharton.upenn.edu

Conflict of Interest: None declared.

Supporting information: Additional supporting information may be found in the online version of this article at the publisher's web site.

R-workspace: The R-workspace `QLMR_files.RData` contains what is needed to reproduce the Table 1 and Figure 3 in the example. The file “`qlmraci_help`” describes the arguments used in the function “`qlmraci`”. (R-supplement.zip)

1. Introduction

Many chronic illnesses, such as schizophrenia, depression, diabetes and cancer, require regular treatment modification that take into account the changing health status of a patient. Long-term management of these illnesses is often made up of a sequence of decision points in which many factors, including response to previous treatments, symptom severity, and side-effects, are considered in deciding if, when, and how, current treatment should be altered. A *dynamic treatment regime* (DTR) operationalizes this sequential decision making and is composed of a sequence of decision rules that take in current patient health information and past treatment as inputs, and output when and how to alter treatment. Dynamic treatment regimes are also known as adaptive interventions [1] or adaptive treatment strategies [2]. The optimal dynamic treatment regime is the regime that optimizes the expected health outcome of interest.

Sequential Multiple Assignment Randomized Trials (SMARTs) [2–5] provide a rich source of longitudinal data that can be used to construct optimal dynamic treatment regimes free from causal confounding. In a SMART, participants move through treatment stages; at each stage patients are possibly re-randomized to different treatments. SMARTs are set apart from traditional randomized trials in that, not only are the participants potentially randomized to sequences of treatments, but also patient response to prior treatment is often included in the randomization scheme of the trial design. For example, it is often the case that individuals who are responding to their current treatment remain on that treatment in the next treatment stage, and only those participants who are considered non-responders to prior treatment are re-randomized at later stages [6–8]. This added flexibility, which allows participants to remain on treatments that are working with the chance for participants to switch away from less successful treatments, in combination with the opportunity to evaluate sequences of treatments, is making SMART studies increasingly common in a variety of health settings. For a list of recently conducted and ongoing SMARTs, see <http://methodology.psu.edu/ra/adap-inter/projects>. We use as a motivating example, the Clinical Antipsychotic Trials of Intervention and Effectiveness (CATIE) study; a well-known SMART study that was designed to compare the effectiveness of sequences of antipsychotic medications for patients with schizophrenia [6, 9].

Q-learning is a regression-based method for constructing optimal dynamic treatment regimes using data collected from a SMART [10–12]. Q-learning is an approximate dynamic programming approach originally developed in the computer science literature and approximates expected outcome, conditional on the available information at each stage, by fitting a regression model. As is the case with any regression based method, the success of Q-learning critically depends on the quality of the regression models fit to the data. However, there has been little attention in the literature so far to this issue of model choice. The authors of [13] consider model checking and diagnostics in the context of the *regret regression* methodology introduced by [14]. The authors of [15] consider model checking with residual diagnostic plots for an alternative estimation method called *g-estimation* [16]. Specifically, Rich et. al. [15] discuss diagnostic tools for g-estimation, which can be used for data collected in observational studies. The three aforementioned methods, i.e., Q-learning, regret regression and g-estimation, are semiparametric and do not require specifying the

distribution of the histories, outcome, residual errors. The authors show that while patterns in residual plots may indicate problems with the model specification, they do not necessarily indicate at which level the problems occur (i.e., the propensity score or the blip function) or how to address them. It may be especially difficult to deal with the situation where more than one of the models is misspecified. Unlike Q-learning, which typically employs linear regression and thus can appeal to a wider audience, the two other approaches (i.e., regret regression and g-estimation) are complex to implement. The issue of model checking for Q-learning has been discussed in [17] where it was shown that the residuals of the standard Q-learning may not be used to detect the omission of important variables that interact with the treatment indicator in predicting the outcome.

We aim to fill this important gap. In particular, we show that owing to the form of the dependent variable used in the regression models in Q-learning, the residuals often suffer from heterogeneity of variance. This heterogeneity is often an artifact of SMART designs in which subsets of individuals achieving “response” to their initial treatments are not re-randomized to an alternative treatment. Moreover, the residual values at each stage may depend on the randomization status at the subsequent stages. We propose a modification of the standard Q-learning approach that avoids these problems and show the residuals obtained by our method can be used to assess the quality of fit in a way analogous to the case of ordinary linear regression. Zhao and Laber [18] proposed to use the information available prior to responder classification to build an estimator of the responder's outcome. However, suggested is not sufficient to produce an interpretable residual plot because it does not solve the issue of observing different patterns in the residual plot of stage 1 Q-learning model based on the response/non-response status.

The remainder of this article is organized as follows. Section 2 introduces data structure and notations, provides an overview of Q-learning, and discusses the problem of variance heterogeneity in the Q-learning context. We introduce our proposed method, Q-learning with mixture residuals (QL-MR), in Section 3. We present the results of simulation studies conducted to examine the performance of our proposed method in Section 4. In Section 5, we apply the proposed method to data collected from CATIE. Finally, we close with a discussion.

2. Data collected from SMARTs

We discuss the general data structure of SMART studies and associated analysis using the Clinical Antipsychotic Trials of Intervention and Effectiveness (CATIE) as an example. CATIE was a multistage clinical trial funded by the National Institute of Mental Health (NIMH) to compare the effectiveness of sequences of antipsychotic medication for the treatment of schizophrenia. Schizophrenia is a chronic illness, and its symptoms include hallucinations, delusions and confused thought and speech processes. Due to side effect burden and variability in treatment response, both between patients and within the same patient over time, effective management of schizophrenia requires regular treatment adaptation. Here, we give only a brief and simplified description of the CATIE study, since the trial has already been described in detail elsewhere [19]. CATIE had two main randomized treatment stages. Two of the treatment options at the initial randomization stage

of randomization were: (1) perphenazine, an older, well-established antipsychotic, and (2) ziprasidone, a newer antipsychotic that was approved by the US Food and Drug Administration around the time the CATIE study began enrollment. Participants who had adequate response to their randomly assigned initial treatment could remain on that medication, while individuals who did not have adequate response were re-randomized to a second-line treatment. We consider two second-line treatment options: (1) olanzapine, which has been shown to be one of the most effective antipsychotics for managing schizophrenia studied in CATIE and (2) one of risperidone or quetiapine; see Figure 1. Below we discuss the generic data structure of two-stage SMARTs, using CATIE as an example.

2.1. Data Structure

In this article, we focus on studies with two stages of treatment decision, as in the CATIE study (Figure 1). A typical SMART data set is composed of n independent and identically distributed trajectories corresponding to n subjects. Each trajectory is composed of a sequence of the form $(O_1, A_1, O_2, S, A_2, Y)$. Here O_j for $j = 1, 2$, consists of a set of pre-treatment covariates at the beginning of the j th stage; that is, the vectors O_1 and O_2 consist of the baseline covariates and intermediate outcomes, respectively. For example in the CATIE study, O_1 includes information gathered on patients prior to initial randomization such as baseline demographics, current and past symptoms, treatments previously taken, and prior medication side effects experienced. Likewise, time-varying measurements such as symptoms, side effects, and medication adherence at clinic visits after this initial visit but before entering the second stage of randomization form O_2 . A_j for $j = 1, 2$, denotes the treatment assigned at j th stage. As noted above, the two stage 1 CATIE treatment options considered here are perphenazine and ziprasidone. Individuals who do not achieve satisfactory results from this first stage treatment option (i.e. classified as non-responders), are re-randomized to an additional second line of treatment. The second line treatments considered here are and one of risperidone or quetiapine. Note that for responders, their initial treatments assigned at stage 1 are merely continued. Let S be a binary variable representing the randomization/responder status at stage 2, coded 1 if an individual is re-randomized at stage 2, and 0 otherwise. Denote the long-term primary outcome of interest, Y . We take the Heinrichs Carpenter QOL scale [20] recorded at 6 months from study enrollment as the outcome of interest. For ease of notation, we define the history at each stage as: $H_1 = O_1$ and $H_2 = (O_1, A_1, O_2)$.

A DTR is formally defined as $d = (d_1, d_2)$ where d_j is a deterministic decision rule that maps covariate information at stage j to an available treatment option at that stage. The optimal DTR, d^{opt} , is the treatment choice that satisfies $\mathbb{E}^{d^{opt}}(Y) \geq \mathbb{E}^d(Y)$, for all d , where $\mathbb{E}^d$ is the expectation when $A_j = d_j(H_j)$.

2.2. Q-learning with Linear Models

Using data collected from a SMART, Q-learning employs approximate dynamic programming to construct optimal dynamic treatment regimes. Q-learning starts from the last stage of the study and finds the treatment option that optimizes the outcome at the final stage. Then, given the optimally-chosen last stage treatment, Q-learning moves backward to

the immediately preceding stage and searches for an treatment option assuming that future treatments will be optimized. The process continues until the first stage is reached. The backward induction procedure followed in Q-learning is designed to avoid treatment options that appear to be optimal in the short term but may lead to a less desirable long-term outcome. While different variations of the standard Q-learning have recently been proposed [21–25], in its simplest and most commonly used form, Q-learning fits a linear regression model for the appropriate dependent variable at each stage.

Without loss of generality, we assume that a higher valued outcome is better. Then the Q-function at stage 2 is defined as: $Q_2(H_2, A_2) = E[Y|H_2, A_2]$, and the stage 1 Q-function is defined $Q_1(H_1, A_1) = E[\max_{a_2} Q_2(H_2, a_2)|H_1, A_1]$. A backward induction argument [26] can be used to prove that the optimal treatment at a particular stage is given by the treatment that maximizes the associated Q-function. In particular, if these two Q-functions were known, the optimal decision rules d_1^{opt} and d_2^{opt} would be generically given by $d_j^{\text{opt}}(h_j) = \arg \max_{a_j} Q_j(h_j, a_j)$, $j = 1, 2$. Thus, $d_2^{\text{opt}}(H_2)$ is the treatment that maximizes the expected value of the primary outcome given the covariate history H_2 ; likewise, $d_1^{\text{opt}}(H_1)$ is the first-stage treatment that, given covariates H_1 , leads to an optimal outcome when combined with the optimal secondstage treatment, $d_2^{\text{opt}}(H_2)$. In practice, however, the true Q-functions are not known and must be estimated from data. Here, as is standard, we consider the following linear working models:

$$Q_j(H_j, A_j; \beta_{j1}, \beta_{j2}) = \beta_{j1}^T H_{j1} + \beta_{j2}^T H_{j2} A_j, \text{ for } j=1, 2,$$

where H_{j1} and H_{j2} are vectors of features from the participant's history at the stage j , H_j . In particular, H_{j2} consists of candidate tailoring variables at the stage j . For example, in CATIE study, H_{22} may include adherence to stage-1 treatment and H_{12} may include baseline demographics and an intercept term.

The traditional Q-learning algorithm can be summarized as follows:

Stage 2:

Estimate the parameters of the stage-2 model by regressing the outcome Y on $(H_{21}, H_{22}A_2)$ among individuals who were re-randomized at stage 2 ($S = 1$) using the model

$$Q_2(H_2, A_2; \beta_{21}, \beta_{22}) = S(\beta_{21}^T H_{21} + \beta_{22}^T H_{22} A_2).$$

This results in estimators $\hat{\beta}_{21}$ and $\hat{\beta}_{22}$.

Stage 1:

1. Construct $\tilde{Y}$, the dependent variable for the stage-1 regression:

$$\tilde{Y} = S \cdot \hat{Y}^{\text{opt}} + (1 - S) \cdot Y,$$

such that $\hat{Y}^{opt}$ is defined as [11, 27] :

$$\hat{Y}^{opt} = \hat{\beta}_{21}^T H_{21} + \max_{a_2} \{ \hat{\beta}_{22}^T H_{22} a_2 \}.$$

In Appendix B of the supplementary materials, we present additional options for constructing $\hat{Y}^{opt}$ that appear in the literature. Note, that while we introduce only one possible version of defining $\hat{Y}^{opt}$ the alternative constructions of $\hat{Y}^{opt}$ also produce $\tilde{Y}$ and residuals that suffer from the same heterogeneity of variance.

2. Estimate the parameters of the stage-1 model by regressing the constructed dependent variable $\tilde{Y}$ on $(H_1, H_{12}A_1)$.

$$Q_1(H_1, A_1; \beta_{11}, \beta_{12}) = \beta_{11}^T H_{11} + \beta_{12}^T H_{12} A_1.$$

This results in the estimates $\hat{\beta}_{11}$ and $\hat{\beta}_{12}$.

Final: Compute optimal dynamic treatment regime using stage-1 and stage-2 estimates:

$$\text{Optimal stage-2 treatment: } \hat{d}_2^{opt}(H_2) = \arg \max_{a_2} Q_2(H_2, a_2; \hat{\beta}_{21}, \hat{\beta}_{22}).$$

$$\text{Optimal stage-1 treatment: } \hat{d}_1^{opt}(H_1) = \arg \max_{a_1} Q_1(H_1, a_1; \hat{\beta}_{11}, \hat{\beta}_{12}).$$

2.3. Checking The Stage-1 Regression Model

Standard regression diagnostic tools can be used to assess the quality of model fit for the stage-2 Q-function. However, non-smoothness in the stage-1 dependent variable causes some problems when trying to apply standard diagnostic tools to models of Q-functions for the first stage. This non-smoothness in $\tilde{Y}$ is caused by the *max* operator present in the definition of $\hat{Y}^{opt}$. Furthermore, the distribution of $\tilde{Y}$ is a mixture distribution, representing two different populations – individuals who are randomized at stage 2 and those who are not. Hence any diagnostics that are based on fitting just one model to the stage-1 dependent variable may be misleading. Specifically, in stage 1 model, we have

$$\mathbb{E}[\tilde{Y} | H_{11}, H_{12}A_1] = \mathbb{E}[S \cdot \hat{Y}^{opt} + (1 - S) \cdot Y] = \beta_{11}^T H_{11} + \beta_{12}^T H_{12} A_1.$$

Let $\varepsilon_1 = S \cdot [\hat{Y}^{opt} - \beta_{11}^T H_{11} - \beta_{12}^T H_{12} A_1]$ and $\varepsilon_2 = (1 - S) \cdot [Y - \beta_{11}^T H_{11} - \beta_{12}^T H_{12} A_1]$.

Regressing $\tilde{Y}$ on (H_1, A_1) implies that $\mathbb{E}[\hat{Y} - \beta_{11}^T H_{11} - \beta_{12}^T H_{12} A_1 | H_{11}, H_{12} A_1] = 0$. However, this does not mean that $\mathbb{E}[\varepsilon_1 | H_1, A_1] = 0$ and $\mathbb{E}[\varepsilon_2 | H_1, A_1] = 0$ that results in observing the different patterns in the residual plot of stage 1 model in standard Q-learning (See Figure 4). Note that, fitting two separate models for $\mathbb{E}[SY^{opt} | H_1, A_1]$ and $\mathbb{E}[(1 - S)Y | H_1, A_1]$ will not alleviate the problem because of the zero-inflated outcomes (i.e., the vector of

SY^{opt} and $(1 - S)Y$ have lots of zeros) which can result in a poor model fit and lead to incorrect confidence intervals and p-values (see Web Appendix C). For example, suppose the outcome is generated from $Y = -A_1O_1 + O_2 + SA_2O_1 + w$, where O_1 , O_2 , O_3 , and w are independent standard Gaussian variables, and $S = \mathbb{I}(O_3 > 1)$. The treatments A_j are Bernoulli random variables with probability 0.5, where A_j s are re-coded as $-1/+1$. Then, the true stage-1 dependent variable is

$$\tilde{Y} = S(-A_1O_1 + O_2 + |O_1|) + (1 - S)(-A_1O_1 + O_2) = -A_1O_1 + O_2 + S|O_1|.$$

Thus, $E[\tilde{Y}|O_1, A_1, S=1] = -A_1O_1 + |O_1|$, and $E[\tilde{Y}|O_1, A_1, S=0] = -A_1O_1$. This toy example shows that the residual plot against O_1 may reveal different patterns depending on the value of S . Web Figure 1 in Appendix C of the supplementary materials visualizes the two patterns in this example (see also Figure 4). The quality of fit of the stage-1 regression may be improved by using more flexible models such as generalized additive models [28]; however, these two different patterns for responders and non-responders will still be present. An additional (albeit related) problem with applying standard residual diagnostic tools to assess the fit of the stage-1 Q-function is variance heterogeneity. Variance heterogeneity in Q-learning is caused by the fact that for some individuals the stage-1 dependent variable is a function of the $\hat{\beta}_2$ and is an expected (i.e. mean) outcome, and for some others it is an observed outcome (see the construction of $\tilde{Y}$ in the stage-1 regression model) [18]. Variance heterogeneity violates one of the standard least squares assumptions.

Because Q-learning involves linear regression, it is tempting to rely on standard residual-based diagnostic plots to diagnose departures from the linear model. Specifically, we focus on plots of residuals against covariates and the histogram of residuals. The former is used to indicate whether the model adequately adjusts for each covariate, or, for example, whether higher order terms need to be added to the model [15]. The histogram of the residuals, on the other hand, is used to check whether the distribution of the random error term is symmetric. For example, if the plot shows bimodality, it may indicate that the distribution of the outcome is a mixture distribution.

3. The Proposed Method: Q-learning with mixture residuals (QL-MR)

In this section we outline our proposed modification to the standard Q-learning approach, which we call Q-learning with mixture residuals (QL-MR), designed to mitigate the problems introduced in the last section. Our proposed approach modifies the residuals of the stage-1 model such that the standard residual plots used to assess regression model fit provide useful guidance for model choice in the Q-learning context as well. More specifically, to avoid variance heterogeneity we fit the stage-2 model among everyone and not just those who were re-randomized at stage 2. We form the stage-1 dependent variable $\tilde{Y}_{QL-MR}$ and estimate the probability of entering the second treatment stage (e.g., being a non-responder) given the observed history up to stage-2 treatment assignment. Then, the stage-1 Q-function is formed using $\tilde{Y}_{QL-MR}$ and decomposed to two parts based on the randomization status at stage 2. For each part, the re-randomization indicator is replaced with its estimate and then a different model is fit for each part. This helps capture possibly

different association patterns of baseline variables H_1 with different components of $\tilde{Y}_{QL-MR}$. Finally, we define the new residuals as a mixture of the residuals obtained from these two models. The proposed stage-1 residuals do not show any trend under correct model specification which accommodates residual analysis using standard tools (see Figure 3). See Web Appendix F for more details. QL-MR can be summarized as follows:

Stage 2:

Estimate the parameters of the stage-2 model by regressing the outcome Y on $(H_{21}, H_{22}A_2)$ using all the individuals, but nesting regression parameters among those with $S = 0$ and $S = 1$, e.g.,

$$Q_2(H_2, A_2; \gamma_{21}, \gamma_{22}, \gamma_{23}) = S(\gamma_{21}^T H_{21} + \gamma_{22}^T H_{22} A_2) + (1 - S)(\gamma_{23}^T H_{23}),$$

where H_{23} is a suitable feature vector for subjects with $S = 0$. This results in the estimates $\hat{\gamma}_{21}$, $\hat{\gamma}_{22}$ and $\hat{\gamma}_{23}$.

Stage 1:

1. Define $\tilde{Y}_{QL-MR}$, the dependent variable for the stage-1 regression:

$$\hat{Y}_{QL-MR} = \max_{a_2} [S(\hat{\gamma}_{21}^T H_{21} + \hat{\gamma}_{22}^T H_{22} a_2) + (1 - S)(\hat{\gamma}_{23}^T H_{23})].$$

Note that unlike the standard Q-learning, the dependent variable in QL-MR is not a mixed function of the estimated parameters and the observed outcomes (see $\tilde{Y}$ in Section 2.2) that solves the variance heterogeneity issue.

2. Postulate a parametric model for $\pi = \Pr(S = 1 | H_2, a)$ and compute the maximum likelihood estimator $\hat{\pi}$. Define $\hat{\pi} = \Pr(S = 1 | H_2, \hat{a})$.
3. Define the stage 1 Q-function $Q_1^*(H_1, A_1, \gamma_1)$, by replacing S with its expected value given H_2 , π .

$$\begin{aligned} Q_1^*(H_1, A_1, \gamma_1) &= \mathbb{E}[\tilde{Y}_{QL-MR} | H_1, A_1] \\ &= \mathbb{E}[\mathbb{E}\{\tilde{Y}_{QL-MR} | H_2\} | H_1, A_1] \\ &= \mathbb{E} \left[\mathbb{E} \left\{ \max_{a_2} [S(\hat{\gamma}_{21}^T H_{21} + \hat{\gamma}_{22}^T H_{22} a_2) + (1 - S)\hat{\gamma}_{23}^T H_{23}] | H_2 \right\} | H_1, A_1 \right] \quad (1) \\ &= \mathbb{E}[\pi \{\hat{\gamma}_{21}^T H_{21} + \hat{\gamma}_{22}^T H_{22}\} | H_1, A_1] + \mathbb{E}[(1 - \pi)\hat{\gamma}_{23}^T H_{23} | H_1, A_1]. \end{aligned}$$

The last equation follows from assuming a contrast coding for the treatment indicator (i.e., $A_1 \in \{-1, +1\}$). Note that, $Q_1^*(H_1, A_1, \gamma_1)$ is a mixture model with two components, namely, responders and non-responders. In the above expression for Q_1^* , we replace π with its estimated value $\hat{\pi}$. Postulate two models for the conditional expectations on the RHS of (1). Assuming linear working models for both conditional expectations on the RHS of (1) possibly with different sets of covariates, we have

$$Q_1^*(H_1, H'_1, A_1, \gamma_1) = [\eta_{11}^T H_{11} + \eta_{12}^T H_{12} A_1] + [\theta_{11}^T H'_{11} + \theta_{12}^T H'_{12} A_1], \quad (2)$$

where $H_1 = (H_{11}, H_{12})$ and $H'_1 = (H'_{11}, H'_{12})$ are the sets of covariates included in the linear models that are not necessarily the same. For example, H_1 may include O_1 while H_2 may include O_1 and O_1^2 . We estimate the parameters η and θ by regressing $\hat{\pi}\{\hat{\gamma}_{21}^T H_{21} + |\hat{\gamma}_{22}^T H_{22}|\}$ and $(1 - \hat{\pi})\hat{\gamma}_{23}^T H_{23}$ on the covariates $(H_{11}, H_{12}A_1)$ and $(H'_{11}, H'_{12}A_1)$, respectively, and denote them by $(\hat{\eta}, \hat{\theta})$.

4. Form residuals $\varepsilon_{\text{QL-MR}}$ as follows

$$\hat{\varepsilon}_{\text{QL-MR}} = \hat{\pi}[\hat{\gamma}_{21}^T H_{21} + |\hat{\gamma}_{22}^T H_{22}|] + (1 - \hat{\pi})\hat{\gamma}_{23}^T H_{23} - [\hat{\eta}_{11}^T H_{11} + \hat{\eta}_{12}^T H_{12} A_1] - [\hat{\theta}_{11}^T H'_{11} + \hat{\theta}_{12}^T H'_{12} A_1]$$

Note that regardless of the value of S (e.g., response/non-response status) every individual have a value of $\hat{\pi}[\hat{\gamma}_{21}^T H_{21} + |\hat{\gamma}_{22}^T H_{22}|]$ and $(1 - \hat{\pi})\hat{\gamma}_{23}^T H_{23}$.

5. Assess residuals $\varepsilon_{\text{QL-MR}}$ using standard residual diagnostics.

If standard residual diagnostics indicate a lack of fit, adjust predictor set accordingly and go back to Stage 1, step 1;

else continue with Q-learning algorithm to construct optimal dynamic treatment regime (Final Step).

Final: Compute optimal dynamic treatment regime using Stage-1 and Stage-2 estimates of final Q-functions:

$$\text{Optimal stage-2 treatment: } d_2^{\text{opt}}(H_2) = \arg \max_{a_2} Q_2(H_2, a_2; \hat{\gamma}_{21}, \hat{\gamma}_{22}).$$

$$\text{Optimal stage-1 treatment: } d_1^{\text{opt}}(H_1) = \arg \max_{a_1} Q_1^*(H_1, a_1; \hat{\gamma}_{11}, \hat{\gamma}_{12}).$$

Remark. After assessing the residual plots based on $\varepsilon_{\text{QL-MR}}$ and selecting required covariates, we can refit the stage-1 model by regressing $\hat{Y}_{\text{QL-MR}}$ on (H_1, A_1) . This improves the efficiency of the parameter estimates in stage-1 model compared to fitting two separate models $\mathbb{E}[\pi\{\hat{\gamma}_{21}^T H_{21} + |\hat{\gamma}_{22}^T H_{22}|\} | H_1, A_1]$ and $\mathbb{E}[(1 - \pi)\hat{\gamma}_{23}^T H_{23} | H_1, A_1]$.

3.1. Asymptotic properties of QL-MR

In this section we study the asymptotic properties of the estimated parameters using QL-MR and compare it with the standard Q-learning. Specifically, we show that while QL-MR addresses the problem of residual heterogeneity, asymptotically it has similar performance as standard Q-learning in terms of estimating the optimal treatment regime. In other words, correction of residual heterogeneity is not at the cost of poor estimation and decision-making.

In stage 2, the estimators $(\hat{\gamma}_{21}, \hat{\gamma}_{22})$ and $(\hat{\beta}_{21}, \hat{\beta}_{22})$ are the same because the extra component $(1 - S)H_{23}$ is orthogonal to (SH_{21}, SH_{22}) . Thus, adding this extra component to the regression model does not change the estimators corresponding to (SH_{21}, SH_{22}) . The parameters of the stage-1 Q-functions in standard Q-learning and QL-MR are asymptotically equivalent under the following two conditions:

C.1 The postulated model for the outcome Y among individuals with $S = 0$ is correctly specified. In other words, $\gamma_{23}^T H_{23}$ is a correctly specified model.

C.2 $\pi = P(S = 1 | H_2, \alpha)$ is correctly specified.

Under conditions C.1 & C.2,

$$\hat{\pi} \{ \hat{\gamma}_{21}^T H_{21} + |\hat{\gamma}_{22}^T H_{22}| \} + (1 - \hat{\pi}) \hat{\gamma}_{23}^T H_{23} = \hat{Y}_{QL-MR} + (1 - S)\tau + o_p(1) = \hat{Y} + o_p(1), \quad \mathbb{E}[\tau | H_2] = 0.$$

where $\tau = Y - \hat{\gamma}_{23}^T H_{23}$ among individuals with $S = 0$. Thus, while asymptotically the stage-1 dependent variables have the same mean in both standard Q-learning and QL-MR, $\tilde{Y}_{QL-MR}$ is less variable than $\tilde{Y}$ for individuals with $S = 0$ [29]. This can alleviate the variance heterogeneity in the residual plots of stage 1 models observed in the standard Q-learning (See Wen Appendix F).

3.2. Confidence intervals for the stage-1 parameters

The asymptotic properties of the stage-2 parameters follow from standard least squares method. However, since the stage-1 dependent variable is a non-smooth function of data, the estimators of stage-1 parameters are nonregular. The authors of [30] propose the Adaptive Confidence Interval (ACI) method which is a generalization of the standard bootstrap to construct confidence intervals for the parameters of stage-1 Q-function in the standard Q-learning. In Appendix A of the supplementary material, we generalize the ACI by replacing the stage-1 dependent veritable with $\hat{\pi} \{ \hat{\gamma}_{21}^T H_{21} + |\hat{\gamma}_{22}^T H_{22}| \} + (1 - \hat{\pi}) \hat{\gamma}_{23}^T H_{23}$. We use the modified version of the ACI to construct valid confidence intervals for the parameters of the stage-1 model in the proposed QL-MR method (see Web Appendix A).

4. Simulation Study

Here we present the results of a simulation study designed to assess the performance of our proposed QL-MR method for using standard residual analyses to select an appropriate model for the stage-1 Q-function. We simulate a two-stage SMART study of 300 individuals using the following generative models. Two continuous-valued baseline variables, O_{11} and O_{12} , are each generated from a standard normal distribution, and the randomized treatments A_j are generated such that $P(A_j = 1) = P(A_j = -1) = 0.5$ for $j = 1, 2$. Finally, intermediate and long-term outcomes are generated as:

$$\begin{aligned} O_{21} &= 5 - 0.3A_1 - 0.5O_{11} + \varepsilon_o, \\ O_{22} &= 5 - 0.3A_1 - 0.5O_{12} + \varepsilon_o, \\ S &= \begin{cases} 1 & \text{if } O_{22} > 5 \\ 0 & \text{otherwise} \end{cases} \\ g(H_2) &= 1 + 2O_{11} - 1.5O_{11}^2 - 2O_{12} + O_{21} - A_1 - 0.5A_1O_{11}, \\ Y &= g(H_2) + S \cdot (0.8O_{21} - 0.5A_2 - 0.4A_2O_{21} - 0.7A_2O_{11}) + \varepsilon, \end{aligned}$$

where $\varepsilon, \varepsilon_o \sim \mathcal{N}(0, 1)$.

We use a linear model for the stage-2 Q-function including all predictors in the appropriate form in order to correctly model the stage-2 Q-functions. Since standard residual analyses can readily be applied to assess and select appropriate stage-2 Q-function, we focus on comparing residual analyses using the stage-1 residuals obtained by our proposed method with the residuals obtained from conventional Q-learning. The stage-1 regression model in our proposed QL-MR method involves three components: (1) $\pi(H_2)$; (2)

$\mathbb{E} \left[\pi \left\{ \hat{\beta}_{21}^T H_{21} + |\hat{\beta}_{22}^T H_{22}| \right\} | H_1, A_1 \right]$; (3) $\mathbb{E} \left[(1 - \pi) \hat{\beta}_{23}^T H_{23} | H_1, A_1 \right]$. We fit a logistic regression model for S to estimate $\pi(H_2)$. As there is a vast literature on how to assess the quality of fit for logistic regression models (see for example, [31–33]), we focus on assessing the fit of the regression models for (2) and (3) in our proposed approach. We fit three increasingly complex linear models using the traditional Q-learning approach, and the QL-MR approach. We start with the following simple regression model that includes just the main effects of O_{11} , O_{12} and A_1 (Model 1):

$$\mathbb{E} \left[\hat{\pi} \left\{ \hat{\beta}_{21}^T H_{21} + |\hat{\beta}_{22}^T H_{22}| \right\} | H_1, A_1 \right] = \beta_{11}H_1 + \beta_{12}A_1, \quad (3)$$

$$\mathbb{E} \left[(1 - \hat{\pi}) \hat{\beta}_{23}^T H_{23} | H_1, A_1 \right] = \beta'_{11}H_1 + \beta'_{12}A_1, \quad (4)$$

where $H_1 = (O_{11}, O_{12})$. We then increase the complexity of the models by including more predictors and more complex functions of predictors, viz., Model 2 including $(O_{11}, O_{12}, A_1, O_{11}^2)$, and Model 3 including $(O_{11}, O_{12}, A_1, O_{11}^2, A_1O_{11})$.

Figure 2 shows the residual plots obtained by fitting Model 1. The first two columns present residual plots against O_{11} and O_{12} , while the third column presents histogram of the stage-1 residuals. The two rows correspond to residuals constructed using standard Q-learning (row 1) and QL-MR (row 2), respectively.

The residual plots against O_{11} shows a “U” trend. So we add the variable O_{11}^2 to H_1 (Model 2). Web Figure 2 shows that this adjustment successfully removes the U-shape trend. However, there is a different trend in the residual plots against O_{11} depending on the initial treatment (the green and orange lines) which indicates the need for adding an interaction term A_1O_{11} (Model 3). Note that interaction terms between A_1 and the baseline variables play a crucial role in tailoring the treatment options. Figure 3 shows that the residual plots

are satisfactory for QL-MR, after these modifications to the predictor set for Q_1^* . However, standard Q-learning still has heterogeneity of variance in the residual plots against O_{11} , even after the same changes. Addition of more terms does not improve the fit. The histogram of the standard Q-learning residuals (*Res QL*) also reveals bimodality. The heterogeneity of variance and the bimodality are caused by the method used to form the stage-1 dependent variable $\tilde{Y}$. More specifically, in case of conventional Q-learning, $\tilde{Y}$ consists of individuals who have taken the optimal stage-2 treatment as well as individuals who have not taken any new treatment at stage 2, which means $\tilde{Y}$ is generated from a mixture density with two components. Figure 4 shows how the residual *Res QL* depends on the value of the variable S . The dashed and solid lines correspond to $S = 0$ and $S = 1$, respectively. This is caused by fitting one model to $\tilde{Y} = S\hat{Y}^{opt} + (1 - S)Y$ which is composed of two components $\hat{Y}^{opt}$ and Y .

In Table 1, we have listed the estimates of the parameters indexing the stage-1 and stage-2 decision rules using standard Q-learning and QL-MR (i.e., (β_{12}, β_{22}) and $(\gamma_{12}, \gamma_{22})$, respectively.). As expected, under conditions C.1 and C.2, the decision rules obtained by both methods are similar. While the residual plots obtained from standard Q-learning fail to adequately check the quality of the model fit, the proposed QL-MR accommodates residual analyses using standard tools.

5. Analysis of CATIE Data

As mentioned in Section 2, CATIE was a practical clinical trial that enrolled patients with schizophrenia. All participants in CATIE were scheduled to follow up with study clinicians monthly. Symptoms, side effects, adherence and a variety of other covariates were collected at the study visits. At baseline, demographics, current and past symptoms, treatment(s) previously taken, and side effect information were gathered on all CATIE participants, who were then randomized to an initial stage-1 treatment. Because CATIE was a practical clinical trial designed to mimic real life, patients could decide, in consultation with their study clinician, to discontinue their current randomized medication if they no longer deemed it adequate (i.e. they did not respond to initial assigned treatment). When it was decided that a CATIE participant was not responding to their initial treatment, they entered a second randomized treatment stage (see Figure 1). Recall, we consider a binary initial randomized treatment to either perphenazine ($A_1 = -1$) or ziprasidone, an atypical antipsychotic ($A_1 = +1$). For stage-2, we consider the binary treatment olanzapine ($A_2 = -1$) versus one of risperidone or quetiapine ($A_2 = +1$). Web Appendix D provides more information about the treatment options in CATIE study.

As with most longitudinal clinical trials, several CATIE participants had missing longitudinal information. As the focus of this paper is on methods for selecting appropriate models for Q-learning, we use a complete CATIE data set that was produced using an imputation strategy described by [34]. Our data set includes 299 patients, 148 of whom were randomized to perphenazine in stage 1 ($A_1 = -1$) and the remaining were assigned to ziprasidone ($A_1 = +1$). Approximately half (166) were categorized as responders to their assigned stage-1 treatment. Of the 133 individuals who did not respond to their stage-1 treatment, 59 (44%) patients were randomized to olanzapine for their stage-2 treatment (A_2

$= -1$) and the remaining participants were randomized to either risperidone or quetiapine ($A_2 = +1$).

Our analysis of the CATIE data set includes baseline symptom severity as measured by the positive and negative syndrome scale [35] (O_{11}), quality of life as measured by the Heinrichs-Carpenter QOL scale [20] (O_{12}) and age at study entry (O_{13}). Intermediate outcomes considered in this analysis were measured prior to stage-2 randomization and include the positive and negative syndrome scale (O_{21}) and medication adherence as measured by the proportion of pills taken in the last month (O_{22}). We take quality of life measured at 6 months as the primary outcome (Y). The Heinrichs-Carpenter QOL scale is a continuous-valued scale with possible scores ranging from 0 to 6 such that a high score implies a higher quality of life. To apply standard Q-learning, we fit the following regression model at stage 2:

$$Y = \beta_{21}^T H_{21} + \beta_{22}^T H_{22},$$

where $H_{21} = (1, O_{11}, O_{11}^2, O_{12}, A_1, O_{21})$ and $H_{22} = (1, A_1)$. Note that unlike the conventional Q-learning that uses only those individuals re-randomized at stage 2 for the stage-2 modeling, QL-MR uses all the individuals to model the stage-2 Q-function as follows:

$$Y = S(\beta_{21}^T H_{21} + \beta_{22}^T H_{22} A_2) + (1 - S)(\beta_{23}^T H_{23}),$$

where $H_{23} = (1, O_{11}, O_{11}^2, O_{12}, A_1, O_{21}) = H_{21}$.

We compare the residuals obtained by our proposed method and conventional Q-learning. Once again, we focus on residual analyses for the stage-1 model and begin with fitting a model which includes just main effects of baseline variables and stage-1 treatment (Model 1). Recall that the proposed method involves three regression models: (1) $\pi(H_2)$; (2)

$\mathbb{E}[\hat{\beta}_{21}^T H_{21} + \hat{\beta}_{22}^T H_{22} | H_1, A_1]$ and (3) $\mathbb{E}[\hat{\beta}_{23}^T H_{23} | H_1, A_1]$. Assuming a logit model for the variable S (responder/non-responder indicator), we fit $\pi(H_2)$ using logistic regression to obtain $\hat{\pi}$, and then postulate the following models:

$$\mathbb{E} \left[\hat{\pi} \left\{ \hat{\beta}_{21}^T H_{21} + \hat{\beta}_{22}^T H_{22} \right\} | H_1, A_1 \right] = \beta_{11} H_1 + \beta_{12} A_1, \quad (5)$$

$$\mathbb{E} \left[(1 - \hat{\pi}) \hat{\beta}_{23}^T H_{23} | H_1, A_1 \right] = \beta'_{11} H_1 + \beta'_{12} A_1, \quad (6)$$

where $H_1 = (O_{11}, O_{12})$.

Figure 5 shows the residual plots obtained by fitting Model 1. The columns correspond to covariates O_{11} , O_{12} and O_{13} . The rows correspond to different methods used to construct the stage-1 dependent variable. The orange and green lines are the loess smoother lines for

individuals with $A_1 = +1$ and $A_1 = -1$, respectively. The figure does not provide strong evidence about existence of interaction between A_1 and baseline variables. However, the residual plot against O_{11} shows a “U” trend. So we add the variable O_{11}^2 to the fitted stage-1 models (Model 2). The residuals of Model 2 are presented in Figure 6. The box plot shows that the variance of the residuals is higher among responders when the stage-1 model is fit using standard Q-learning but it is more homogeneous when QL-MR is used. Also, the residual standard errors based on the proposed method are almost half of the standard Q-learning (see also Web Figure 3).

Table 2 reports the parameter estimates and 90% confidence limits for the parameters of Model 2 where Stage-1 model includes all the main effects and O_{11}^2 (O_{13} is omitted since it was not significant). Both standard Q-learning and the proposed QL-MR result in roughly similar parameter estimates and confidence intervals. Note that for compatibility, only $(\gamma_{21}, \gamma_{22})$ are reported as the stage-2 parameters of QL-MR. Due to the non-regularity of the stage-1 regression parameters we have used ACI and the modified version of ACI to construct valid confidence intervals for the stage-1 parameters of standard Q-learning and QL-MR, respectively. For more information, see Appendix A of the supplementary material and [22, 36], [30], [37] and [38]. Our analysis shows that the mean outcome difference between those who were assigned resperidone/quetiapine vs. olanzapine at stage 2 and ziprasidone ($A_1 = +1$) at stage 1 (i.e., $\mathbb{E} Y|O_{11}, O_{12}, A_1 = +1, O_{21}, A_2 = +1] - \mathbb{E} Y|O_{11}, O_{12}, A_1 = +1, O_{21}, A_2 = -1]$) is -0.30 with 90% CI (-0.60, 0.004). This means these patients may benefit from olanzapine as stage-2 treatment. However, the mean outcome difference between stage-2 treatment options given initial treatment with perphenazine ($A_1 = -1$) does not show any evidence in the favor of a particular stage-2 treatment (point estimate of 0.06 with 90% C.I. (-0.19, 0.33)). Our analysis does not show any significant difference between stage-1 treatments as the treatment effect estimate is very close to zero and is not statistically significant (point estimate of -0.01 with 90% C.I. (-0.10, 0.11)).

6. Discussion and Concluding Remarks

Checking the adequacy of model fit is important for all model-based inference; yet, until now there has been little discussion of model assessment in the existing dynamic treatment regimes literature. In this paper, we proposed a novel modification to traditional Q-learning for constructing optimal dynamic treatment regimes; the proposed residuals can be used along with standard diagnostic tools for checking the adequacy of model fit for the Q-functions. Similar to many model based approach, the validity of Q-learning is compromised when inaccurate models are fit to the data; thus the proposed approach will improve the quality and scope of applying Q-learning in practice. This is particularly relevant in SMART studies where often only a subgroup of participants (e.g. non-responders to the initial treatment) are re-randomized at stage 2.

We believe that future research utilizing our proposed residual will broaden the reach and applicability of Q-learning for constructing dynamic treatment regimes. Specifically, detecting high-leverage points, outliers and/or influential points is important in Q-learning just as in standard regression analysis. Traditionally, modified versions of regression residuals are used to identify these observations; in particular, one often uses studentized

residuals, defined as $\hat{\varepsilon}/\sqrt{\text{Var}(\hat{\varepsilon})}$, for the purpose of detecting high-leverage points in regression analysis. In future work, we will consider how to compute an analogous quantity using our newly proposed QL-MR method. Extending studentized residuals to our proposed residuals, $\hat{\varepsilon}_{\text{QL-MR}}$, is not straightforward as it requires estimation of $\text{Var}(\hat{\varepsilon}_{\text{QL-MR}})$. Deriving an analytical expression for the variance may at best be messy, but some carefully crafted resampling approach, e.g. m -out-of- n bootstrap [36] or ACI [30], may potentially be useful. Detecting outliers in case of linear regression models typically employs jackknife residuals; the i th jackknife residual is defined as $(y_i - \hat{y}_{(i)})/\sqrt{\text{Var}(y_i - \hat{y}_{(i)})}$, where y_i is the i th observed outcome and $\hat{y}_{(i)}$ is its predicted value based on the model built without the i th observation (as in leave-one-out cross-validation). While it is not hard to imagine a jackknife version of our newly proposed residual in the Q-learning context, estimation of the corresponding variance remains the key to its success. Likewise, computing Cook's distance for detecting influential points in a Q-learning model faces the same issue. Addressing these open questions will expand the scope and applicability of Q-learning.

Laber et. al. [29] proposed a version of Q-learning called interactive Q-learning that replaces modeling the predicted future optimal outcomes $\tilde{Y}$ (i.e., the stage-1 dependent variable in standard Q-learning) with two mean-variance function modelings [39]. This improves the model interpretability and can handle the possible non-linear dependence between $\tilde{Y}$ and baseline variables. Adaptation of our proposed method to interactive Q-learning can potentially be used to provide a residual analysis for this particular type of Q-learning method.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Dr. Ertefaie is supported by award number P50 DA010075 and SES-1260782, from the National Institute on Drug Abuse and National Science Foundation. Dr. Chakraborty's research is partially supported by the startup grant from the Duke-NUS Graduate Medical School, Singapore.

References

1. Collins L, Murphy S, Bierman K. A conceptual framework for adaptive preventive interventions. *Prevention Science*. 2004; 5(3):185–196. [PubMed: 15470938]
2. Lavori P, 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.
3. Thall P, Millikan R, Sung H, et al. Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine*. 2000; 19(8):1011–1028. [PubMed: 10790677]
4. Murphy S. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*. 2005; 24(10):1455–1481. [PubMed: 15586395]
5. Murphy S, Lynch K, Oslin D, McKay J, TenHave T. Developing adaptive treatment strategies in substance abuse research. *Drug and alcohol dependence*. 2007; 88:S24–S30. [PubMed: 17056207]
6. Stroup TS, McEvoy JP, Swartz MS, Byerly MJ, Glick ID, Canive JM, McGee MF, Simpson GM, Stevens MC, Lieberman JA. The national institute of mental health clinical antipsychotic trials of

- intervention effectiveness (catie) project. *Schizophrenia Bulletin*. 2003; 29(1):15–31. [PubMed: 12908658]
7. Kasari C. Developmental and augmented intervention for facilitating expressive language (ccnia). 2012; 26 [ClinicalTrials.gov](https://www.clinicaltrials.gov) database, updated Apr 2009.
 8. Lei H, Nahum-Shani I, Lynch K, Oslin D, Murphy S. *Annual Review of Clinical Psychology*. 2012; 8:21–48.
 9. Swartz MS, Perkins DO, Stroup TS, McEvoy JP, Nieri JM, Haak DC. Assessing clinical and functional outcomes in the clinical antipsychotic trials of intervention effectiveness (catie) schizophrenia trial. *Schizophrenia Bulletin*. 2003; 29(1):33–43. [PubMed: 12908659]
 10. Watkins, C. PhD Thesis. King's College; Cambridge: 1989. Learning from delayed rewards.
 11. Murphy SA. A generalization error for q-learning. *Journal of machine learning research: JMLR*. 2005; 6:1073–1097. [PubMed: 16763665]
 12. Schulte PJ, Tsatis AA, Laber EB, Davidian M. Q-and a-learning methods for estimating optimal dynamic treatment regimes. *arXiv preprint arXiv:1202.4177*. 2012
 13. Henderson R, Ansell P, Alshibani D. Regret-regression for optimal dynamic treatment regimes. *Biometrics*. 2010; 66(4):1192–1201. [PubMed: 20002404]
 14. Murphy S. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2003; 65(2):331–355.
 15. 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):1–24.
 16. Robins, J. *Proceedings of the Second Seattle Symposium on Biostatistics*. Springer; New York: 2004. Optimal structural nested models for optimal sequential decisions; p. 189–326.
 17. Chakraborty, B.; Moodie, EE. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Vol. 76. Springer Science & Business Media; 2013.
 18. Zhao YQ, Laber EB. Estimation of optimal dynamic treatment regimes. *Clinical Trials*. 2014; 11(4):400–407. [PubMed: 24872361]
 19. Stroup TS, McEvoy JP, Swartz MS, Byerly MJ, Glick ID, Canive JM, McGee M, Simpson GM, Stevens MD, Lieberman JA. The National Institute of Mental Health clinical antipsychotic trials of intervention effectiveness (CATIE) project: schizophrenia trial design and protocol development. 2003; 29(1):15–31.
 20. Heinrichs DW, Hanlon TE, Carpenter WT. Quality of Life Scale: An instrument for rating the schizophrenic deficit syndrome. 1984; 10:388–398.
 21. Zhao Y, Kosorok M, Zeng D. Reinforcement learning design for cancer clinical trials. *Statistics in medicine*. 2009; 28(26):3294–3315. [PubMed: 19750510]
 22. Chakraborty B, Murphy S, Strecher V. Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*. 2010; 19(3):317–343. [PubMed: 19608604]
 23. Shortreed SM, Laber E, Lizotte DJ, Stroup TS, Pineau J, Murphy SA. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine learning*. 2011; 84(1-2):109–136. [PubMed: 21799585]
 24. Song R, Wang W, Zeng D, Kosorok MR. Penalized q-learning for dynamic treatment regimes. *Statistica Sinica*. 2012; 10.5705/ss.2012.364
 25. Goldberg Y, Kosorok MR. Q-learning with censored data. *Annals of statistics*. 2012; 40(1):529–560. [PubMed: 22754029]
 26. Sutton, RS.; Barto, AG. *Reinforcement learning: An introduction*. The MIT Press; 1998.
 27. Nahum-Shani I, Qian M, Almirall D, Pelham WE, Gnagy B, Fabiano GA, Waxmonsky JG, Yu J, Murphy SA. Q-learning: A data analysis method for constructing adaptive interventions. *Psychological methods*. 2012; 17(4):478–494. [PubMed: 23025434]
 28. Hastie T, Tibshirani R. Generalized additive models. *Statistical science*. 1986:297–310.
 29. Laber EB, Linn KA, Stefanski LA. Interactive model building for q-learning. *Biometrika*. 2014:asu043.

30. Laber EB, Lizotte DJ, Qian M, Pelham WE, Murphy SA. Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*. 2014; 8(1):1225–1272. [PubMed: 25356091]
31. Hosmer, DW., Jr; Lemeshow, S.; Sturdivant, RX. *Applied logistic regression*. John Wiley & Sons; 2004.
32. Neter, J.; Wasserman, W.; Kutner, MH., et al. *Applied linear statistical models*. McGraw-Hill; 1996.
33. Pregibon D. Logistic regression diagnostics. *The Annals of Statistics*. 1981; 9(4):705–724.
34. Shortreed SM, Laber E, Scott Stroup T, Pineau J, Murphy SA. A multiple imputation strategy for sequential multiple assignment randomized trials. *Statistics in medicine*. 2014; 33(24):4202–4214. [PubMed: 24919867]
35. Kay SR, Flazbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. 1987; 13(2):261–76.
36. Chakraborty B, Laber E, Zhao Y. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*. 2013; 69(3):714–723. [PubMed: 23845276]
37. Moodie EE, Richardson TS. Estimating optimal dynamic regimes: Correcting bias under the null. *Scandinavian Journal of Statistics*. 2010; 37(1):126–146.
38. Laber E, Murphy S. Adaptive confidence intervals for the test error in classification. *Journal of the American Statistical Association*. 2011; 106(495):904–913. [PubMed: 22053123]
39. Carroll, RJ.; Ruppert, D. *Transformation and weighting in regression*. Vol. 30. CRC Press; 1988.

Figure 1.
SMART design of the CATIE study.

Figure 2. Simulation: Model 1 residual plots against baseline variables O_{11} and O_{12} and the residuals histogram. The rows are residuals constructed based on Y^{opt} , and the proposed QL-MR method, respectively. The orange and green lines are the loess smoother lines for individuals with $A_1 = +1$ and $A_1 = -1$, respectively.

Figure 3. Simulation: Model 3 residual plots against baseline variables O_{11} and O_{12} and the residuals histogram. The rows are residuals constructed based on Y^{opt} , and the proposed QL-MR method, respectively. The orange and green lines are the loess smoother lines for individuals with $A_1 = +1$ and $A_1 = -1$, respectively.

Figure 4. Simulation: Residual plots against baseline variables O_{11} . The residuals are constructed based on Y^{opt} . The first and the second columns are the residuals obtained in scenario 1 & 2, respectively. The dashed and solid lines correspond to individuals with $S = 1$ and $S = 0$, respectively.

Figure 5.

CATIE: Model 1 residual plots against baseline variables O_{11} , O_{12} and O_{13} . The rows are residuals constructed based on Y^{opt} and the proposed QL-MR method, respectively. The orange and green lines are the loess smoother lines for individuals with $A_1 = +1$ and $A_1 = -1$, respectively.

Figure 6. CATIE: Model 2 residual plots against baseline variables O_{11} , O_{12} and O_{13} . The first rows are residuals constructed based on Y^{opt} and the proposed QL-MR method, respectively. The orange and green lines are the loess smoother lines for individuals with $A_1 = +1$ and $A_1 = -1$, respectively. The last row is the box plots of the residuals given the response(resp)/non-response(non-resp) status.

Table 1

Simulated data: Estimates of the Stage-2 and Stage-1 decision rule parameters.

Parameter	Standard Q-learning			QL-MR		
	Estimate	90% CI		Estimate	90% CI	
Stage-2 Model						
A_2	-2.00	-2.61	-1.38	-2.00	-2.64	-1.44
A_2O_{11}	-1.48	-1.63	-1.34	-1.48	-1.62	-1.35
A_2O_{21}	1.61	1.49	1.73	1.61	1.49	1.74
Stage-1 Model						
A_1	-1.02	-1.46	-0.58	-1.05	-1.50	-0.66
A_1O_{11}	-3.41	-3.81	-2.98	-3.47	-3.88	-3.08

CATIE: Stage-2 and Stage-1 regression models. In QL-MR, the stage-2 parameters correspond to $(\gamma_{21}, \gamma_{22})$.

Table 2

Parameter	Standard Q-learning			QL-MR		
	Estimate	90% C.I		Estimate	90% C.I	
Stage-2 Model						
O_{11} : Baseline PANSS	0.01	-0.12	0.14	0.02	-0.11	0.15
O_{11}^2 : Baseline PANSS	0.05	-0.02	0.13	0.02	-0.05	0.09
O_{12} : Baseline Quality of Life	0.48	0.36	0.61	0.49	0.37	0.60
A_1 : Stage-1 treatment	0.004	-0.11	0.12	0.008	-0.09	0.11
O_{21} : PANSS during stage-1	-0.19	-0.30	-0.08	-0.20	-0.30	-0.10
A_2 : Stage-2 treatment	-0.06	-0.17	0.04	-0.07	-0.17	0.03
A_2A_1	-0.09	-0.19	0.02	-0.09	-0.19	0.01
Stage-1 Model						
O_{11} : Baseline PANSS	-0.13	-0.23	-0.04	-0.12	-0.22	-0.03
O_{11}^2 : Baseline PANSS	0.06	0.00	0.12	0.05	-0.01	0.12
O_{12} : Baseline Quality of Life	0.51	0.42	0.61	0.50	0.42	0.59
A_1 : Stage-1 treatment	-0.01	-0.10	0.11	-0.01	-0.13	0.09