Mach Learn (2011) 84:109–136
DOI 10.1007/s10994-010-5229-0

Informing sequential clinical decision-making through reinforcement learning: an empirical study

Susan M. Shortreed · Eric Laber · Daniel J. Lizotte · T. Scott Stroup · Joelle Pineau · Susan A. Murphy

Received: 26 February 2010 / Revised: 23 September 2010 / Accepted: 14 November 2010 /
Published online: 22 December 2010

Abstract This paper highlights the role that reinforcement learning can play in the optimization of treatment policies for chronic illnesses. Before applying any off-the-shelf reinforcement learning methods in this setting, we must first tackle a number of challenges. We outline some of these challenges and present methods for overcoming them. First, we describe a multiple imputation approach to overcome the problem of missing data. Second, we discuss the use of function approximation in the context of a highly variable observation set. Finally, we discuss approaches to summarizing the evidence in the data for recommending a particular action and quantifying the uncertainty around the Q-function of the recommended policy. We present the results of applying these methods to real clinical trial data of patients with schizophrenia.

Keywords Optimal treatment policies · Fitted Q-iteration · Policy uncertainty

Editors: S. Whiteson and M. Littman.

S.M. Shortreed (✉) · J. Pineau
School of Computer Science, McGill University, Montreal, QC, H3A 2T5, Canada
e-mail: shortreed.s@ghc.org

J. Pineau
e-mail: jpineau@cs.mcgill.ca

T.S. Stroup
NYS Psychiatric Institute, Room 2703 Unit/Box:100, 1051 Riverside Drive, New York, NY 10032, USA
e-mail: stroups@pi.cpmc.columbia.edu

E. Laber · D.J. Lizotte · S.A. Murphy
Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

E. Laber
e-mail: laber@umich.edu

D.J. Lizotte
e-mail: danjl@umich.edu

S.A. Murphy
e-mail: samurphy@umich.edu

1 Introduction

There has been a surge of interest in recent years within the medical community regarding the application of Reinforcement Learning (RL) techniques to optimize treatment policies (Thall and Wathan 2000; Ernst et al. 2006; Murphy et al. 2007; Guez et al. 2008; Zhao et al. 2009). Recent efforts have targeted the design of treatment policies for cancer, epilepsy, depression, and HIV/AIDS, among others. The opportunities for the RL field to have a deep societal impact in this area are substantial.

The RL literature provides a number of off-the-shelf methods for automatic optimization of action sequences; yet, many challenges remain before RL-based policies can be deployed to inform clinical decision making. In particular, we have identified four key challenges:

1. In many cases, training data is collected during randomized trials, with little or no opportunity to control exploration or to acquire further data to evaluate policies (policy evaluation on new data would involve running a new trial, which can be a massive undertaking).
2. As clinical data is extremely expensive to obtain, both in terms of time and money, the number of trajectories available to us is modest compared to what is available in simulation or most traditional RL settings—typically much fewer than 5000 trajectories are collected from any given trial. In addition, the observation space of most clinical trials is typically continuous, highly variable, and high-dimensional.
3. The few trajectories we do have will frequently have missing observations, and we often do not have a good understanding of the reasons for this missingness (or partial observability).
4. Before any recommended policy can be accepted by the medical community, it is imperative to quantify the evidence for, and the uncertainty around, the recommended policy. Measures of uncertainty are essential at all decision making levels, ranging from high level policy makers to the clinicians providing care and guidance to patients.

This paper provides an in-depth case study of using RL to optimize treatment choices for people with schizophrenia based on real data from a two-stage clinical trial involving 1460 patients. Throughout the methodology section, we propose and discuss methods for overcoming the key technical challenges outlined above. We show how these can be used in the RL setting to handle real clinical data. In particular, we discuss the use of multiple imputation to overcome the missing data problem. We then present two methods for quantifying the evidence in the data for the choices made by the learned optimal policy. These methods are designed to use only the same training data that was available to learn the optimal policy. The first method, called bootstrap voting, can be used to visually convey the evidence for the *action choices* made by the learned policy. From our experience, this method may be a particularly useful tool for conveying the evidence supporting an action choice to scientists and clinicians. The second method uses recently developed methodology to provide confidence intervals for the learned Q-values. This gives us a formal and rigorous measure of whether the values of two actions are significantly different.

We begin by reviewing the RL framework for use with clinical data in Sect. 2. Section 3 describes the data set at the core of our case-study, along with some of the particular challenges of applying RL methods to this type of data. In Sect. 4, we introduce the core methodologies used to overcome missing data, learn a policy, and present the evidence for a particular action choice and uncertainty around an action's Q-function. The results of applying these methods to a clinical trial of treatments for schizophrenic patients are presented in Sect. 5. Finally, in Sect. 6, we conclude with a discussion of the results and methods,

detailing some of the limitations of these and other RL methods when applied to clinical trial data and indicating areas for further research.

2 Reinforcement learning of treatment policies: notation and models

The Reinforcement Learning methodology provides a framework for an agent to interact with its environment and receive rewards based on observed states and actions taken, with the goal of learning the best sequence of actions to maximize its expected sum of rewards (Sutton and Barto 1998). This problem design is similar to the problem of estimating treatment policies in the medical sciences (Murphy 2003; Pineau et al. 2007). Chronic illnesses like diabetes, epilepsy, depression, HIV/AIDS, among others, require multi-stage decision making. At each stage, treatment is necessarily adapted to a patient's response, as defined by symptom severity, treatment adherence, changes in side-effect severity, developed drug-resistance, evolution of co-existing medical conditions, and a multitude of other factors. Thus, the clinician is faced with the task of observing a patient's history (state) and recommending a treatment (action) that maximizes the patient's long-term clinical outcome (cumulative reward). Clinicians wanting to construct principled (i.e. evidence based) rules for tailoring treatment have begun to run sequential multiple assignment randomized trials (SMART) (Murphy 2005; Dawson and Lavori 2004; Rush et al. 2004; Thall and Wathan 2000).

The data from a SMART study consists of a sequence of treatments together with a sequence of observations for each patient in the trial. Such data are essentially trajectories of control inputs (treatments or actions) and outputs (observations or rewards), and fit naturally into the reinforcement learning framework. Analysis using RL methods enables us to recover not just an optimal *treatment* but an optimal *sequence of treatments* (i.e. a policy) that is tailored to each individual patient. The use of SMART studies in the medical community provides substantial opportunities for the use of RL methods to automatically learn sequences of treatments that optimize expected patient outcomes.

We now describe RL methods as they pertain to data collected from a SMART study. In the following we use upper case letters to denote random variables, such as S and A, and lower case letters, such as s and a, to denote realizations or observed values of the random variables. We also use $B, K, J, M,$ and T as integer constants, this distinction between integer constants and random variables should be clear from the context. We assume a finite horizon Markov decision process (MDP) with a total of T stages. The action (treatment) at stage t is denoted by $A_t \in \mathcal{A}_t$, and the state by S_t collected prior to treatment assignment. We assume a pre-defined reward observed after each action, $R_t = R(S_t, A_t, S_{t+1})$, that is a function of the state at stage t, the action taken at stage t, and the resulting state at stage $t+1$. Thus, the data at stage t, $t = 1{:}T$ is comprised of the triplet (S_t, A_t, R_t). The goal is to use RL methods to determine the sequence of actions (treatment assignments) that will maximize the expected total reward over the study period,

$$\mathbb{E}\left[\sum_{t=1}^{T} R_t\right]. \tag{1}$$

In a SMART study, the treatment actions are sampled according to a fixed (known) random exploration policy. This random policy allows us to simply use the state information measured and actions made to compute an unbiased estimate of the effect of an action on the total reward. Each of the n individuals in a trial contributes one data trajectory:

$(S_{1,i}, A_{1,i}, R_{1,i}, S_{2,i}, A_{2,i}, R_{2,i}, \ldots, S_{T,i}, A_{T,i}, R_{T,i})$. In most SMART studies to date, the number of stages, T, is small, usually between two to four stages.

We use the standard Bellman equations to define the optimal state-action value function at stage t. The Q-function is defined as

$$Q_t(s_t, a_t) = \mathbb{E}\left[R_t + \max_{a_{t+1} \in \mathcal{A}_{t+1}} Q_{t+1}(S_{t+1}, a_{t+1}) \,\middle|\, S_t = s_t, A_t = a_t\right], \tag{2}$$

with $Q_T(s_T, a_T) = R_T$. The optimal action at stage t, denoted $\pi_t^*(s_t)$, is defined as the action that maximizes the state-action value function, Q_t:

$$\pi_t^*(s_t) = \arg\max_{a_t \in \mathcal{A}_t} Q_t(s_t, a_t). \tag{3}$$

We use batch off-policy fitted Q-iteration to learn the optimal treatment policy (Ernst et al. 2005). The optimal action in the last stage of the trial, T, is found by estimating $\text{argmax}_{a_T} \mathbb{E}[R_T(S_T, A_T, S_{t+1})|S_t = s_t, A_t = a_t]$. The optimal actions at earlier stages are then estimated by rolling the estimated optimal value functions into the earlier state-action value functions.

3 CATIE, a SMART study

The Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) was an 18 month multistage clinical trial of 1460 patients with schizophrenia. This study was funded by the National Institute of Mental Health in order to evaluate the clinical effectiveness of sequences of specific antipsychotic drugs. As this study is described in detail elsewhere (Stroup et al. 2003; Swartz et al. 2003), we give a simplified overview of the CATIE study here. CATIE is a SMART study with two major treatment stages ($T = 2$). At entry into the study, participants were randomized to one of five stage 1 treatments: olanzapine, risperidone, quetiapine, ziprasidone, and perphenazine. Patients were then followed up for 18 months and allowed to switch treatment if their assigned stage 1 treatment was not effective. The result of this protocol is the time duration for each of the two treatment stages in CATIE is patient dependent. For example, some patients did well on their initial treatment and thus remained in stage 1, hence the time duration of stage 1 for these patients is 18 months.

Treatment discontinuation in the first stage was generally due to either a lack of *efficacy* of the prescribed treatment (i.e. symptoms remained high), or a lack of *tolerability* of the treatment (i.e. side-effects were substantial). If failure of the stage 1 treatment occurred, i.e. a patient chose to discontinue their stage 1 treatment, patients were given the choice to enter one of two arms in stage 2: the efficacy arm or the tolerability arm. Patients who chose the efficacy arm were re-randomized to clozapine (50%) or one of olanzapine, risperidone or quetiapine (16.6% each). Patients who chose the tolerability arm were re-randomized to olanzapine, risperidone, quetiapine, or ziprasidone (25% each). We note here that clozapine can have dangerous side-effects, requiring that patients taking clozapine submit to regular blood tests and screenings. As a result of a desire to avoid these risks and extra monitoring, some individuals who discontinued their previous treatment due to lack of efficacy chose to enter the tolerability arm of stage 2 where clozapine was not an option. The time duration for stage 2 is 18 minus the month in which the patient entered stage 2.

3.1 Reward definition and state representation

Throughout the 18 month period, patients were scheduled for monthly visits with their medical practitioner and detailed data was recorded every three months. This collected information includes approximately twenty demographic variables, recorded at baseline, and thirty variables measured repeatedly throughout the study. These variables include symptom levels, quality of life, side-effect burden, and medication compliance, among others. From these variables, we must extract both a reward function and state representation for each of the two decision stages.

Throughout this paper, we take as the primary outcome the Positive and Negative Syndrome Scale (PANSS) score (Kay et al. 1987). The PANSS score is a medical scale designed to measure symptom severity in patients with schizophrenia. A low PANSS score indicates few psychotic symptoms; as an individual's psychotic symptoms increase, so does their PANSS score. We would like to learn the optimal sequence of treatments that minimizes a patient's schizophrenic symptoms, as measured by the PANSS score, over the course of the entire 18 months of the CATIE study. While not a topic of this paper, it is worth pointing out that a variety of other functions of symptom measurement or entirely different outcomes could be considered (e.g. side-effects, treatment cost, etc.). Many open questions remain as to how to best combine multiple (sometimes competing) outcomes into a single scalar reward, as required by the traditional RL framework. We define the reward function as the negative of the area under the PANSS score curve over the 18 month trial, denoted $\mathrm{AUC}_{\mathrm{PANSS}}(18)$. We note here that a patient's PANSS score fluctuates over time (see Sect. 3.3, Fig. 2), even under treatment, and once a patient's PANSS score is low, it does not mean that it will remain low. Using a composite total reward, such as $\mathrm{AUC}_{\mathrm{PANSS}}(18)$, ensures that the optimal treatment will be one that not only reduces symptoms quickly, but also helps maintain a low PANSS score throughout the study.

Defining the state representation for medical applications of RL techniques is a non-trivial problem; there are often many variables measured on each patient both at the beginning of the study and throughout the study, of which many have continuous domains. Since determining the appropriate state representation is not a main topic of this paper, we give a brief summary and motivation for the state representation we implement. To begin, we differentiate between two types of state variables. The first type of state variable is thought to be potentially useful for making decisions concerning treatment, and thus should be used to inform the learned policy. The second type of state variable may not be useful for making treatment decisions per se but may be predictive of future rewards nonetheless.

We include the PANSS score, measured at entry into a treatment stage, in this first type of variable. This score is the primary method of symptom measurement in patients with schizophrenia; it is monitored closely in practice, and we expect it to be useful in informing treatment decisions. Regarding the second type of variable, it is often the case that there are many variables that can aid in the accurate estimation of the Q-function, but that are not needed to inform the policy because their influence on the Q-function is the same for all actions. For example, a patient's age may be helpful for predicting his or her overall outcome, but it may not be useful for selecting an appropriate treatment action. By including these variables as part of state space, we can bring our domain representation closer to satisfying the Markov assumption, that is, the assumption that the reward and state transition distributions depend only on current state.

We identified five of these second type of variables: a binary indicator for the presence of the side effect tardive dyskinesia at entry into the CATIE study, a binary variable indicating whether a patient was hospitalized for a psychotic episode in the three months

prior to enrollment into the CATIE study, a categorical variable indicating what type of site the patient was treated in (private practice, state mental health clinic, university clinic, VA, or a multi-function site), the length of time in the CATIE study prior to the current stage, and previous treatment. The first three variables are state variables for both stages 1 and 2, whereas the last two variables are included in the state representation only for stage 2.

3.2 Missing data

Schizophrenia is a chronic disease, characterized by abnormalities in a person's perception of reality. Symptoms include hallucinations, delusions, and confused speech and thought processes. The illness places a high level of burden on individuals afflicted with the disease, as well as those who care for them. In studies of antipsychotics in patients with schizophrenia, dropout is often high, between 30 and 60% (Adams 2002). CATIE was designed to be an 18 month longitudinal study with 1460 participants randomized to stage 1 treatment. Of these, 705 patients completed the 18 months of the study, while the remaining 755 (51.7%) dropped out before study completion. While this dropout is the primary source of missing data, there are a number of instances of *item missingness* in this particular data set as well. Item missingness occurs when patients miss a monthly visit (but show up to later visits), or do not provide all information requested. The pattern of missing data is similar in all time-varying variables collected during the CATIE study, thus we visually display the missing data pattern for our variable of interest, PANSS, in Fig. 1 as an example. From this figure, we can see that the proportion of missing data in CATIE due to participant dropout is high, whereas the amount of item missingness is nominal.

Clearly, patient dropout reduces the amount of available data and thus increases the *variance* of the estimates of Q-values and action choices. What may not be immediately apparent, however, is that dropout can also introduce bias into our learned action values and policies. When data is missing in trials, it is tempting to simply remove individuals who have missing data and perform an analysis on the remaining individuals; this is

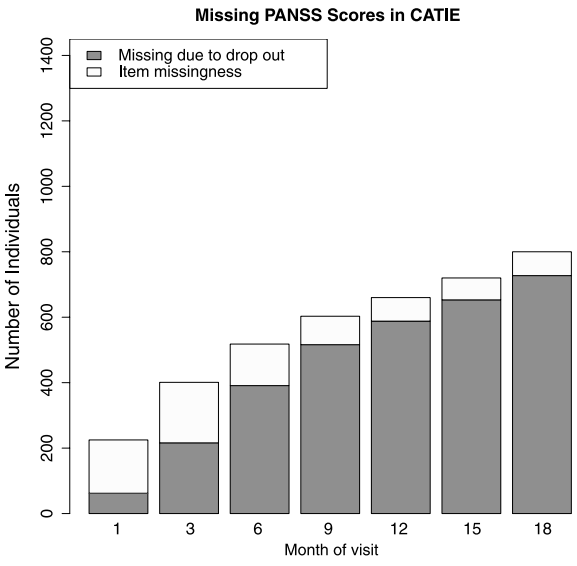

Fig. 1 Barplot of missing PANSS scores in the CATIE study. The total height of the bar shows the absolute number of people who have a missing PANSS score at each of these monthly visits. The *dark grey* area represents the number of people who have missing PANSS score because they dropped out of the study prior to that month. The *unshaded area* is the number of missing PANSS scores due to item missingness. The missing data pattern for other time-varying patient information collected during the CATIE study is similar to the missing data pattern shown here

often termed a *complete case analysis* in the statistics literature. Unfortunately, in many situations, a complete case analysis will lead to biased estimates of the treatment effect. This bias can occur in any data collection procedure, but is especially common in medical trials (Little and Rubin 1987; NAP 2010). In our analysis of the CATIE data, described in Sect. 4.1 below, we apply multiple imputation (Little and Rubin 1987; Rubin 1996; Schafer 1999) to use all the available data (individuals with both complete and incomplete data) in order to reduce bias.

3.3 Variability and small sample size

As has already been discussed, data collection in clinical trials is very expensive, both in time and money, and because of this, large sample sizes are rare. A recent overview of studies involving patients with schizophrenia revealed that the average number of people in such trials was 62 (a median of 60). A mere 58 out of 1,941 trials had sample sizes larger than 300 (Adams 2002). In light of this, with a sample size of 1460, we can consider CATIE a large randomized trial, especially in the field of schizophrenia research.

There are two other factors, besides training set size, that complicate the estimation of the value function from this type of data. First, the type of information collected on patients can be extremely variable, both across patients, and within a patient over time. The variability in the PANSS score is illustrated in Fig. 2. Figure 2(a) displays all observed PANSS scores at each exam using boxplots. The length of the interquartile ranges and the presence of outliers illustrates the high variability in the PANSS scores in CATIE. Figure 2(b) shows the PANSS scores for 100 randomly selected CATIE participants over the course of the trial; each line represents an individual's observed PANSS scores by month of observation. This graph shows the fluctuating nature of symptoms in patients with schizophrenia.

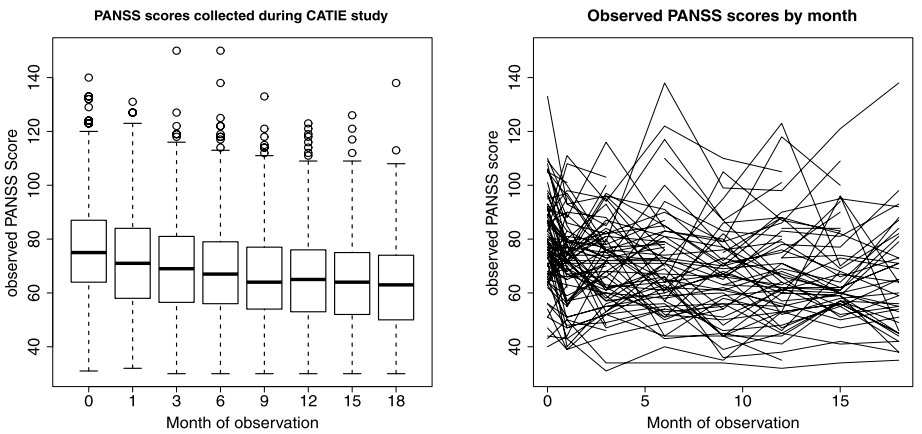

Fig. 2 (**a**) Boxplot of observed PANSS score collected during the CATIE study. The *boxes* represent the inter-quartile range of the observed PANSS scores at each of these months, the *line bisecting the boxes* is the median score, and the *circles* represent outlying scores. (**b**) Line graph representing the observed PANSS scores of 100 randomly selected CATIE participants by month of observation

A more subtle concern in fitting the Q-function, is that the treatment effects we are trying to detect are often relatively small; that is most actions are *expected to perform nearly equally well a priori*. The main reason for this is because clinical trials must satisfy basic ethical requirements, including ensuring that patients are always getting the best quality of care and treatment possible, given our current medical knowledge. For this reason, it is unethical to include treatments which are a priori known to be worse than others. This suggests that the measured differences in Q-values between the various treatments are likely to be small, which only amplifies the problems of variable data and small training set sizes.

4 Methods

In this section, we detail the methods employed to address the above challenges in using clinical data to learn sequences of actions and to characterize our uncertainty in these action choices. The goal of the analysis is to learn the optimal treatment policy for minimizing the expected area under the PANSS curve over the 18 months of the study. We use the methods presented in this section to learn the optimal treatment policy and present the evidence in the training set for our learned policy. In Sect. 4.1 we discuss multiple imputation methods that can be used to overcome missing data. Then in Sect. 4.2 we review linear function approximation for fitting Q-functions to batch data. In Sect. 4.3.1, we introduce bootstrap voting, an exploratory data analysis method that can help to illustrate the evidence for a recommended action. In Sect. 4.3.2, we describe a recently developed method for estimating confidence intervals around the coefficients used to measure the effect of an action on the value of the Q-function.

4.1 Overcoming missing data

As highlighted above, a significant portion of our data set is missing, either due to patient dropout or item missingness. This problem is not without precedent in the reinforcement learning literature, where the Partially Observable Markov Decision Process (POMDP) paradigm was developed to handle cases in which the observation does not fully reveal the state of the system (Smallwood and Sondik 1973; Monahan 1982; Kaelbling et al. 1998). However, the POMDP model makes specific assumptions about the patterns of missingness (i.e. which variables represent latent states and which variables represent observations). In contrast, in a clinical trial data set, almost all variables have missing values for one or more time points and all variables are observed at some time points. For example, patient i may have an observed PANSS score at months 0, 1, 2, 3, 5, 6 of the 18 month study, whereas patient j may have observed PANSS score at months 0, 1, 3, 6, 8, 9, 12 of the 18 month study. In short, if the main reason for considering latent variables is to deal with missingness, then there is no clear distinction between which variables should be considered latent and which should be considered observed.

A common method for handling missing data in the statistics literature is multiple imputation (Little and Rubin 1987; Rubin 1996; Schafer 1999). As in the POMDP setting, one builds and learns a model for the complete (e.g. including missing or latent variables) data. However, instead of using the model (as in POMDPs) to learn a policy, multiple imputation only uses the model to fill in the missing values. The missing values are filled in repeatedly to create a collection of imputed training data sets, with observed values left unchanged.

One can then apply standard (fully observable) RL algorithms to learn the Q-function for each of the training data sets. Parameter estimates, $\hat{\theta}$, are combined across all imputations in order to obtain one estimate for the parameter of interest via (4) (Little and Rubin 1987).

$$\hat{\theta}^* = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_M \approx \mathbb{E}_{\mathcal{D}_{\text{miss}}|\mathcal{D}_{\text{obs}}}(\hat{\theta}) = \int \hat{\theta}\ P(\mathcal{D}_{\text{miss}}|\mathcal{D}_{\text{obs}})d\mathcal{D}_{\text{miss}}, \tag{4}$$

where M is the number of imputed data sets.

Bayesian regression techniques are often used to estimate the imputation models. One specifies a model for the complete data (see below for how we model the complete data), assumes a prior distribution over the unknown model parameters, and estimates the posterior of the parameters using the observed data. Missing values are then replaced with samples from the posterior predictive distribution of the missing data given the observed. This is done by plugging in parameter values sampled from the posterior distribution and a sampled random error value into the complete data model (Gelman et al. 1995).

While multiple imputation methodology is very general, in practice, various structural assumptions are usually made (depending on the properties of the domain), to preserve computational tractability and ensure better stability in cases where the observed data is limited. In the analysis of the CATIE study, we use two techniques for estimating the complete data distribution and performing the multiple imputations. We first use fully conditional specification (FCS) (van Buuren et al. 2006; van Buuren 2007) to estimate all missing variables. Two benefits of FCS are that it scales reasonably well with the number of variables and it allows the flexibility to easily use different models for each type of variable, including binary, categorical and continuous variables. However, it is difficult to implement a model that enforces smoothness over time in the mean of a time-varying variable such as PANSS in the FCS framework. Since we expect the mean of the symptom measure PANSS to be smooth across time, we use a second method, based on a mixed effects model (Schafer 1997; Schafer and Yucel 2002), to re-impute individual monthly PANSS scores according to the desired smoothness constraint. Note that we only use the PANSS scores imputed through the mixed effect model for Q-learning. The PANSS scores imputed through FCS are discarded; however, it is desirable to include PANSS in the FCS procedure to allow better imputation of the other variables.

Overview of the FCS method Denote all of the variables in the data set $\mathcal{D}$ by $v_0, v_1, v_2, \ldots, v_J$; in the imputation process no distinction between state variables, rewards, and actions is necessary. We order these variables by month, for variables collected at the same month we order the variables from the fewest number of missing values to the most. We use v_0 to denote the set of baseline variables that contain no missing information, v_1 to denote the variable collected at baseline with the least amount (but at least some) missing information, with v_J representing the variable collected at month 18 with the greatest number of missing values. The model for the complete data in FCS is formed via conditional models for each variable v_j given $v_0, \ldots, v_{j-1}$. The parameters in the conditional models are not shared across models, and their priors are independent. The crucial assumption underlying the FCS imputation is that these conditional models are able to predict missing values of the variable v_j, despite the fact that the model was built using only those individuals for whom v_j was observed. In order to ensure the validity of this assumption, we have included as many predictors as possible leading to a rich imputation model. Under this assumption we can estimate the jth conditional model using only the patients with observed v_j.

Let $v_{j_{\text{miss}}}$ denote the missing observations in variable v_j and $v_{j_{\text{obs}}}$ the observed values in variable v_j. The estimation of each conditional model is interweaved with the imputations as follows. First, we estimate a model for the variable v_1 using v_0, the patients with observed v_1 and the prior. Then we sample from the posterior predictive model for $v_{1_{\text{miss}}}|v_0, v_{1_{\text{obs}}}$ and fill in missing values of v_1. Then using these imputed values, we estimate the conditional model for the variable v_2 given v_1, v_0 and the prior using the patients with observed v_2. All of these patients have a value for v_1 due to the previous imputation. We sample from $v_{2_{\text{miss}}}|v_0, v_1, v_{2_{\text{obs}}}$ to impute the missing values of v_2, and so on. Each of these conditional models is estimated using regression techniques chosen to suit the type of v_j (i.e. integer-valued, real-valued, etc.) and the types of the predictors. Using this formulation, the posterior predictive distribution of the missing data, given the observed, is:

$$P(\mathcal{D}_{\text{miss}}|\mathcal{D}_{\text{obs}}) = \prod_{j=1}^{J} P(v_{j_{\text{miss}}}|v_0, v_1, v_2, \ldots, v_{j-1}, v_{j_{\text{obs}}}),$$

where

$$P(v_{j_{\text{miss}}}|v_0, v_1, v_2, \ldots, v_{j-1}, v_{j_{\text{obs}}}) = \frac{\int P(v_j|v_0, v_2, \ldots, v_{j-1}, \theta_j)\pi_j(\theta_j)\,d\theta_j}{\iint P(v_j|v_0, v_1, v_2, \ldots, v_{j-1}, \theta_j)\pi_j(\theta_j)\,d\theta_j\,dv_{j_{\text{miss}}}},$$

$P(v_j|v_0, v_1, v_2, \ldots, v_{j-1}, \theta_j)$ denotes the conditional distribution, and π_j denotes the prior on θ_j.

Overview of the Bayesian mixed effects method In the following, we use a subscript i to denote patient i, and the subscript m to denote month. In this method, we model the PANSS score at each month, m, given treatment, prior PANSS scores, and other predictors collected both prior to, and at month m, via a mixed effects model. The mixed effects model includes a random intercept term, $g_i \sim \mathcal{N}(0, \sigma_g)$, for each patient, a random error term, $\epsilon_{m,i} \sim \mathcal{N}(0, \sigma_\epsilon)$, for each (patient, month) pair, and a piecewise linear regression spline on the month of observation (Hastie et al. 2001), with knots, ξ, at month 1 and continuing at monthly intervals until month 17. The random intercept term, g_i, models the correlation between the PANSS symptom scores at different months within an individual (Diggle et al. 2002). Since the PANSS scores vary independently across individuals, the random intercept terms, g_i, $i = 1, \ldots, 1460$, are independent. The error terms, $\epsilon_{m,i}$, $i = 1, \ldots, 1460$, $m = 1, \ldots 18$, are also independent. A PANSS score measured for individual i at month m is modeled by

$$(\gamma_0 + g_i) + \underset{\sim}{\gamma}^{\mathsf{T}} \tilde{s}_{m,i} + \eta a_{m,i} + \sum_{\xi=1}^{17} \nu_\xi (m - \xi)_+ + \epsilon_{m,i}, \tag{5}$$

where $a_{m,i}$ denotes the ith patient's treatment at month m, $\tilde{s}_{m,i}$ denotes the vector of predictors including PANSS collected prior to month m as well as other variables collected both prior to and at month m, and the ν_ξ's are the coefficients for the spline and are constrained such that (5) is continuous in m. We fit this model using diffuse priors on the coefficients using the patients who have observed PANSS at month m, and assume that the model holds for individuals missing PANSS scores at month m. From this model, we sample from the posterior predictive distribution to impute the missing PANSS values in $\mathcal{D}$. Algorithm 1 summarizes the full imputation procedure.

Algorithm 1 Algorithm for imputing missing data in CATIE. Note: $v_0, v_1, \ldots, v_J$ denote the variables in $\mathcal{D}$, ordered by month; for variables collected at the same month, we order the variables from the fewest number of missing values to the most. We use v_o to denote the set of baseline variables that contain no missing information.

Input: A training data set $\mathcal{D}$ that has missing values.
First, impute all variables:
for $j = 1{:}J$ **do**
Specify the conditional model $v_j|v_0, v_1, \ldots, v_{j-1}, \theta_j$
Using the prior and the conditional model, estimate the posterior distribution of the parameters in the model for $v_{j_{\text{miss}}}|v_0, v_1, \ldots, v_{j-1}, v_{j_{\text{obs}}}$.
Sample values for the parameters in the conditional model from the estimated posterior distributions.
for each patient i in $\mathcal{D}$ who is missing v_j **do**
Fill in missing value $v_{j,i}$ with a draw from the posterior predictive distribution of $v_{j_{\text{miss}}}|v_1, \ldots, v_{j-1}, v_{j_{\text{obs}}}$, by plugging the parameter values from the posterior distribution into the conditional model and sampling a random error term from the appropriate distribution.
end for
end for
Second, re-impute all missing PANSS variables:
Use observed PANSS score, specified multivariate distribution of PANSS across the months, with mean PANSS score for individual i at month m equal to $(\gamma_0 + g_i) + \underset{\sim}{\gamma}^{\mathsf{T}} \tilde{s}_{m,i} + \eta a_{m,i} + \sum_{\xi=1}^{17} \nu_\xi (m - \xi)_+$, and a diffuse prior over the parameters $(\sigma_g, \sigma_\epsilon, \underset{\sim}{\gamma}, \eta, \nu_1, \nu_2, \ldots, \nu_{17})$ to estimate the posterior distribution over the parameters.
Sample values for the parameters in the conditional model from the estimated posterior distributions.
for $m = 1{:}18$ **do**
for each patient i in $\mathcal{D}$ who is missing PANSS at month m **do**
Fill in PANSS at month m with samples from the posterior predictive distribution by sampling $g_i \sim \mathcal{N}(0, \sigma_g)$ and $\epsilon_i \sim \mathcal{N}(0, \epsilon)$, substituting the parameter values from above along with $a_{i,m}$, and $\tilde{s}_{im}$, into the above formula.
end for
end for
return A single completed CATIE data set in which all missing values are replaced with imputations drawn from the posterior predictive distribution.

4.2 Smoothing the Q-function for highly variable data

The next algorithmic challenge is to learn the optimal state-action value function. Since we are working with a training data set of patient trajectories (or to be more specific, imputed training sets of patient trajectories), we focus on the fitted Q-iteration approach. The core idea of fitted Q-iteration is to learn the Q-function via regression. Many regression models can be considered. Given that data from clinical trials are often highly variable and contain few trajectories, we favor a simple linear regression function. That is, we use linear function approximation to estimate the state-action value function (Lagoudakis and Parr 2003; Irodova and Sloan 2005; Ernst et al. 2005; Parr et al. 2008). Suppose there are T stages: For $t = 1, \ldots, T$, let $\mathbb{1}_{A_t=k}$ be the indicator function that action A_t is equal to action k with a 1 if true and 0 otherwise. Define the feature vector, $x(s_t, a_t)$, to be a function of the state and action at stage t. Let $x(s_t, a_t)$ be of length $d_t \cdot |\mathcal{A}_t|$, where d_t is the number of state variables at stage t:

$$x(s_t, a_t) = [s_t^{\mathsf{T}} \mathbb{1}_{a_t=1}, s_t^{\mathsf{T}} \mathbb{1}_{a_t=2}, \ldots, s_t^{\mathsf{T}} \mathbb{1}_{a_t=|\mathcal{A}_t|}]^{\mathsf{T}}. \quad (6)$$

We model the state-action value via

$$Q_t(s_t, a_t; \beta_t) = \beta_t^{\mathsf{T}} x(s_t, a_t) = \sum_{k=1}^{|\mathcal{A}_t|} \beta_{t,k}^{\mathsf{T}} s_t \mathbb{1}_{a_t=k}. \tag{7}$$

The least-squares estimator at the last stage, $\hat{\beta}_T$, minimizes the sum of squared errors between the Q-value at time T, $Q_T = R_T$, and the fitted Q-function, $\hat{Q}_T$, over the training data (Lagoudakis and Parr 2003; Parr et al. 2008):

$$\begin{aligned}\hat{\beta}_T &= \operatorname{argmin} \sum_{i=1}^{n} \left(R_{T,i} - Q_T(s_{T,i}, a_{T,i}; \beta_T)\right)^2 \\ &= \operatorname{argmin} \sum_{i=1}^{n} \left(R_{T,i} - \beta_T^{\mathsf{T}} x(s_{T,i}, a_{T,i})\right)^2.\end{aligned}$$

The least squares estimators for earlier time intervals, $(T-1):1$, are

$$\hat{\beta}_t = \operatorname{argmin} \sum_{i=1}^{n} \left(\tilde{Q}_{t,i} - Q_t(s_{t,i}, a_{t,i}; \beta_t)\right)^2,$$

where $\tilde{Q}_{t,i} = R_{t,i} + \max_{a_{t+1} \in \mathcal{A}_t} \hat{Q}_{t+1}(s_{t+1,i}, a_{t+1})$. Thus the least squares estimator for β_t is:

$$\hat{\beta}_t = \left(\sum_{i=1}^{n} x(s_{t,i}, a_{t,i}) x(s_{t,i}, a_{t,i})^{\mathsf{T}}\right)^{-1} \sum_{i=1}^{n} x(s_{t,i}, a_{t,i}) \tilde{Q}_{t,i}. \tag{8}$$

As discussed in Sect. 3.1, some state variables (tardive dyskinesia, recent psychotic episode, clinic site type, etc.) are included in the state representation to aid in accurately estimating the Q-function but are not used to inform treatment. We implement this in the model for the Q-function by constraining the influence of these variables on the Q-function to be equal across all actions. That is, we include one term for each of these variables as opposed to multiple terms (i.e., one per each treatment action). Different forms of modeling the Q-function (i.e. non-linear) may necessitate more complex ways of enforcing this constraint on the parameters. Including variables in this way can improve the quality of our Q-value estimates without adding too many parameters to our model.

4.3 Illustrating and quantifying uncertainty in learned policies

In addition to learning the optimal treatment policy, we need to convey the evidence for this learned policy and our uncertainty in the estimated Q-functions for this policy. Uncertainty arises due to small training set size and the variability in the data—we cannot be certain that our learned optimal action choices are correct. This section outlines two novel methods: *bootstrap voting*, for conveying the evidence for action choices and *adaptive confidence intervals*, for measuring uncertainty about the expected performance of this policy.

Quantifying uncertainty is important in the medical field, as it is crucial to know *if* we have enough evidence to recommend one treatment over another. Especially when the training set size is small, it is helpful to know *if we do not have enough evidence to recommend one and only one treatment*. In such a situation, the best decision may be to report back to

clinicians which treatments are candidates for optimal treatments and which treatments we can conclude are inferior given the training set collected. Measures of uncertainty can also help to ensure that we do not produce a treatment policy that is needlessly complex. Some patient variables may not be very useful in determining which treatment is best. If we can identify those variables that are helpful and those that are not, we can reduce the burden of information collection for health care providers.

4.3.1 Bootstrap voting

Bootstrap voting is an exploratory data analysis method that graphically presents the evidence in the data for a learned action choice at a given stage (Lizotte et al. 2009). Furthermore, it illustrates how the evidence for the optimal learned action changes as a function of state. In this section, we review the methods for computing bootstrap vote estimates for each action. Since the methodology described below can be implemented in the same manner at each stage and for each state, we drop references to both stage and state. We simply consider K possible actions, and use Q to denote the Q-function value given a specific stage and state.

Consider the case where we would like to compare K actions. Suppose we do this by running $\binom{K}{2}$ "head-to-head" trials comparing each pair of actions. Each action would be involved in $K-1$ of these trials. The parameter of interest in bootstrap voting is the probability that action k would "win", i.e. appear to be optimal, in each of the $K-1$ trials in which it participates.[1] We denote this quantity p_k^{win}. Note that for $K > 2$ we may not have $\sum_{k=1}^{K} p_k^{win} = 1$, since there is some probability that the set of $\binom{K}{2}$ pairwise trials will produce a "discordant" result (i.e. intransitivity among pairwise results). The more closely matched the actions, the smaller the training set sizes, and the more variable the data, the higher the probability we have of observing a discordant result. We denote the probability of a discordant result $p_\emptyset^{win}$.

Formally, we will estimate, p_k^{win}, defined as

$$p_k^{win} = \prod_{k';k\neq k'}^{K} p_{k,k'} = \prod_{k';k\neq k'}^{K} P_{\mathcal{D}}(\hat{Q}(a=k) > \hat{Q}(a=k')) \tag{9}$$

$$\text{with } p_\emptyset^{win} = 1 - \sum_{k=1}^{K} p_k^{win}.$$

The probabilities in (9) are over the distribution of training sets generated by future sets of pairwise trials. As discussed above, running future trials is extremely difficult and expensive, thus we use the training set we already have collected in order to estimate p_k^{win} for each action. We do this by using the bootstrap, a resampling method which is often employed to estimate properties of complex statistics (Efron 1979; Efron and Tibshirani 1993). A bootstrap sample, $\mathcal{D}^b$, of training set $\mathcal{D}$ of size n, is obtained by sampling n trajectories with replacement from the original training set. The bootstrap is a popular method because the

[1] One may think that a more natural quantity would be the probability that each action "wins" in a future K-arm trial. However, it can be shown that this quantity is affected by correlations between the Q estimates in a way that can induce a non-intuitive ranking of the treatments. In particular, an action may look worse according to this measure even if it has the highest Q-value and the variances of Q estimates are all equal. In these settings, p^{win} maintains the natural ordering of actions and also provides information about uncertainty in the estimates.

variability we observe in the resampled datasets can be a good approximation to the variability we would observe if we were to draw datasets from the true data generating distribution. Thus, we can use bootstrap samples to gain valuable insight into the variability of learned parameters from training set to training set.

The simplest way to estimate p_k^{win} using the bootstrap is to first generate B re-sampled training sets and calculate $\hat{Q}^b(\cdot)$ for each one. We can then define the estimates:

$$\hat{p}_{k,k'} = (1/B)\sum_{b=1}^{B} \mathbb{1}[\hat{Q}^b(a=k) > \hat{Q}^b(a=k')],$$

where $\mathbb{1}[\cdot]$ is the indicator function, returning 1 for true and 0 for false. We then compute $\hat{p}_k^{win} = \prod_{k':k\neq k'}^{K} \hat{p}_{k,k'}$. In practice, this estimator can have very high variance from training set to training set, especially in cases where there is in fact no difference in the true Q-values. For example, in the case where $K = 2$ and both actions have equal Q-values: $Q(a=1) = Q(a=2)$; the true probability $p_1^{win} = p_2^{win} = 0.5$. However, in this situation the above estimate $\hat{p}_1^{win}$ will have a uniform distribution between 0 and 1 from training set to training set. This means it is very likely for the above estimator to give the impression that there is a significant difference in Q-values when in fact there may not be. In order to produce an estimator with lower variance, we use the double bootstrap (Efron and Tibshirani 1993). A double bootstrap sample, $\mathcal{D}^{b,b}$, is a random sample of n individuals from the bootstrapped sample $\mathcal{D}^b$. Using the double bootstrap samples we can compute $\hat{Q}^{b,b'}(a=k)$ and $\hat{Q}^{b,b'}(a=k')$ and compare them. We can then use a two-level average of these comparisons over the double bootstrap samples to form our estimates $\hat{p}_{k,k'}$:

$$\begin{aligned}\hat{p}^b_{k,k'} &= (1/B)\sum_{b'=1}^{B} \mathbb{1}[\hat{Q}^{b,b'}(a=k) > \hat{Q}^{b,b'}(a=k')],\\ \hat{p}_{k,k'} &= (1/B)\sum_{b=1}^{B} \hat{p}^b_{k,k'}.\end{aligned} \tag{10}$$

We then compute $\hat{p}_k^{win} = \prod_{k':k\neq k'}^{K} \hat{p}_{k,k'}$ as before.

This procedure produces an estimator with lower variance, while introducing some bias towards there being no difference between actions—it is effectively a bagged version of the first estimator (Breiman 1996). It is in a sense "conservative", that is, it assumes that there is a smaller difference between the Q-values of different actions than does the non-bagged estimator and is less likely to give the spurious results described above. The main drawback of the double-bootstrap-based estimator is that it is extremely computationally intensive, especially in our application given that for each double bootstrap sample, we must generate multiple imputations as described in Sect. 4.1. In order to help ease the computational burden, we use an approximation to the double bootstrap estimator (Lizotte et al. 2009). We replace our double bootstrap estimate of $\hat{p}^b_{k,k'}$ with

$$\hat{p}^{b*}_{k,k'} = \Phi\left(\frac{\hat{Q}^b(a=k) - \hat{Q}^b(a=k')}{\mathrm{se}_{k,k'}}\right),$$

where Φ is the cumulative distribution function for a Gaussian distribution with mean $= 0$ and variance $= 1$, and $\mathrm{se}_{k,k'}$ corresponds to the standard error of the estimate $\hat{Q}^b(a=k) - \hat{Q}^b(a=k')$. We use the bootstrap to estimate $\mathrm{se}_{k,k'}$ as in (11) below. To compute $\hat{p}_{k,k'}$, we use (10), replacing $\hat{p}^b_{k,k'}$ with $\hat{p}^{b*}_{k,k'}$. Algorithmic details of our implementation of the

Algorithm 2 Double bootstrap approximation algorithm for estimating an array of $p_k^{\text{win}}(s_t)$ values, the probability that action k will win in all pairwise comparisons against all other actions at time interval t for state value s_t.

Input: Training set $\mathcal{D}$; B the number of bootstrap iterations; M the number of imputations; T vectors $\{\mathcal{S}_t\}$ of state values; T sets $\mathcal{A}_t$, listing all possible actions for each t.
Using each of B bootstrapped training sets, first learn functions $\hat{Q}_t^b$ parameterized by $\hat{\beta}_t^b$ $\forall t$:
for $b = 1$ **to** B **do**
  Draw bootstrap sample $\mathcal{D}^b$ by sampling trajectories from $\mathcal{D}$ with replacement.
  for $m = 1$ **to** M **do**
    Impute missing data in $\mathcal{D}^b$ using Algorithm 1, giving completed training sets $\mathcal{D}^{b,m}$.
    Apply batch fitted Q-iteration to $\mathcal{D}^{b,m}$ giving estimated parameters $\hat{\beta}_t^{b,m}$ of $\hat{Q}_t^{b,m}$ $\forall t$.
  end for
  Compute bootstrap estimate $\hat{\beta}_t^b$ as average over the $\hat{\beta}_t^{b,m}$:

$$\hat{\beta}_t^b = \frac{1}{M}\sum_{m=1}^{M} \hat{\beta}_t^{b,m} \quad t \in \{1, \ldots, T\}.$$

end for
for $t \in \{1, \ldots, T\}$, $s_t \in \mathcal{S}_t$, $k \in \mathcal{A}_t$, $k' \in \mathcal{A}_t \setminus k$ **do**
  Define $\hat{\Delta}_{k,k'}^b(s_t) \triangleq \hat{Q}_t^b(a_t = k, s_t) - \hat{Q}_t^b(a_t = k', s_t)$.
  Calculate sample mean of the bootstrap action differences

$$\bar{\Delta}_{k,k'}(s_t) = \frac{1}{B}\sum_{b=1}^{B} \hat{\Delta}_{k,k'}^b(s_t).$$

  Compute the bootstrap estimate for standard error of action difference using the sample standard deviation of the $\Delta_{k,k'}^b(s_t)$:

$$\hat{\text{se}}_{k,k'}(s_t) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\Delta}_{k,k'}^b(s_t) - \bar{\Delta}_{k,k'}(s_t))^2}. \tag{11}$$

  Calculate the double bootstrap approximation $\hat{p}_{k,k'}^{b*}$ using

$$\hat{p}_{k,k'}^{b*} = \Phi\left(\frac{\hat{\Delta}_{k,k'}^b(s_t)}{\hat{\text{se}}_{k,k'}(s_t)}\right) \equiv \Phi\left(\frac{\hat{Q}_t^b(a_t = k, s_t) - \hat{Q}_t^b(a_t = k', s_t)}{\hat{\text{se}}_{k,k'}(s_t)},\right)$$

  where Φ is the cumulative density function for the standard Gaussian distribution.
end for
Estimate $\hat{p}_{k,k'}(x)$ as average of the B bootstrap estimates

$$\hat{p}_{k,k'}(s_t) = \frac{1}{B}\sum_{b=1}^{B} \hat{p}_{k,k'}^{b*}(s_t).$$

Calculate

$$\hat{p}_k^{\text{win}}(s_t) = \prod_{k \in \mathcal{A}_t \setminus k'} \hat{p}_{k,k'}(s_t).$$

return An array, $\hat{p}_k^{\text{win}}(s_t)$, of estimates for each action at each time interval t, and for all state values s_t.

bootstrap voting procedure are given in Algorithm 2. The bootstrap voting methodology is useful, as we will see in Sect. 5, for illustrating the evidence for the recommended policy to clinicians in a way that is interpretable. It is especially helpful for conveying the evidence

for choices among more than two actions; however, it should not be interpreted as a formal measure of confidence for the effect of the policy.

4.3.2 Adaptive confidence intervals

In the clinical trials setting, confidence intervals are commonly used for providing rigorous statements about the strength of evidence in the training data for, or against, a particular scientific hypothesis. For example:

- Is there sufficient evidence in the training data to conclude that the treatment action recommended by the learned optimal policy is significantly better than competing actions?
- Is there insufficient evidence for a single best action, and what subset (if any) of the actions can be excluded as non-optimal?
- Is a particular state variable necessary for making the above decisions?

In summary, the purpose of the confidence interval is to communicate to clinicians the degree of evidence in the training data for a particular treatment or for the usefulness of a patient variable in decision making.

In most RL applications, confidence intervals play a very different role. For example, in (Kaelbling et al. 1996), (Strehl and Littman 2004), and (Strehl and Littman 2005), measures of confidence are used to guide exploration. Their bounds are usually very conservative, and produce large confidence intervals. Since the most important property of the confidence interval in that setting is that it successfully pinpoints which action should be tried next, the exact location of the ends of the confidence intervals are less important; more important is the relative ordering of actions conveyed by the confidence intervals. In our case, where the goal of the confidence intervals is to communicate the quality of the learned policy to an outside audience, not only are the length of the confidence interval and location of the end points important, but it is also essential that we can specify the fraction of the time the confidence interval contains the true Q-values (e.g. if the confidence level is 95%, then 95% of the training sets should lead to a confidence interval that covers the true Q-value).

Another line of research in RL, related to confidence intervals, is the estimation of variance, for example estimating the variance of the value function of a particular policy (Mannor et al. 2007; Tetreault et al. 2007). These methods are restricted to discrete state spaces, and do not apply readily to our problem. Furthermore, the clinical trial setting is somewhat different as we have one (small) training set, thus we are forced to use this training set to learn the policy and then reuse the same training set to provide measures of confidence concerning this policy. Resampling methods such as the bootstrap are often used in the clinical trial setting to provide both approximate variance estimates and confidence intervals (Efron 1979; Efron and Tibshirani 1993).

However, there are difficulties in extending standard statistical approaches, such as the bootstrap, to the Q-learning setting. In particular, because maximization is used in learning the policy and maximization is a non-differentiable function, extra care must be taken in constructing a confidence interval. It is known (Shao 1994) that commonly employed results like the consistency of the bootstrap distribution and the Central Limit Theorem do not hold in cases where the true parameters—in this case the true Q-values—are at or near these points of non-differentiability. The max operator used in the Bellman equation is differentiable at a given state s when there is a unique optimal action for that state. Conversely, the max operator is non-differentiable when, given a state s, the values of two or more actions are equal and optimal. When the training set is sufficiently small so that it is difficult to

discriminate between an optimal and other nearly optimal actions, adjustments to resampling methods like the bootstrap are required to produce confidence intervals that possess the desired confidence level (Andrews 2000).

In the CATIE study, there are only two stages, thus only stage 1 estimators are impacted by the non-differentiable maximization. We employ a bootstrap resampling method developed in Laber et al. to construct confidence intervals for the stage 1 action-value function in the CATIE study. To aid in the exposition, we outline the method of Laber et al. for a two-stage SMART study with two possible actions ($a_j = 1$ or 2) at each stage. However, the ideas and results presented here generalize readily to an arbitrary finite number of stages and treatments (see Laber et al. 2010).

Recall that the model for the stage t Q-function is $Q_t(s_t, a_t; \beta_t) = x(s_t, a_t)^{\mathsf{T}}\beta_t$. Given a state s and action k, we want to construct a confidence interval for the stage 1 Q-function: $Q_1(s, k; \beta_1) = x(s, k)^{\mathsf{T}}\beta_1$. We begin by using a hypothesis test at each realized state, s_2 to decide if multiple stage 2 actions produce an equivalent value. Define the test statistic:

$$\mathcal{T}(s_2) = \frac{n}{\log(n)} \frac{(\hat{\beta}_2^{\mathsf{T}}(x(s_2, 1) - x(s_2, 2)))^2}{(x(s_2, 1) - x(s_2, 2))^{\mathsf{T}}\hat{\Sigma}_2(x(s_2, 1) - x(s_2, 2))}$$

$$\text{where } \hat{\Sigma}_2 \triangleq \frac{1}{n}\sum_{i=1}^{n} x(s_{2,i}, a_{2,i})x(s_{2,i}, a_{2,i})^{\mathsf{T}}$$

and n is the size of the training data set. This test statistic is a measure of a standardized difference between values for treatment actions 1 and 2 at state s_2. With the exception of the $\log(n)$ term, this test statistic resembles an F statistic in regression (Neter et al. 1996). We interpret $\mathcal{T}(s_2) > 1$ as there being strong evidence that the values of the actions differ in state s_2; conversely $\mathcal{T}(s_2) \leq 1$ indicates that the actions may be equivalent. Consider three sets of states:

1. The set of states for which the $\mathcal{T}(s_2) > 1$; we conclude that for these states the best treatment action is unique.
2. The set of states for which the $\mathcal{T}(s_2) \leq 1$ and in *truth* $Q_2(s_2, 1) - Q_2(s_2, 2) = 0$; these are states in which the best treatment is not unique.
3. The set of states for which the $\mathcal{T}(s_2) \leq 1$ and in *truth* $Q_2(s_2, 1) - Q_2(s_2, 2) \neq 0$; these are states in which it appears that the best treatment may not be unique, when in fact it is unique.

Of course we could only differentiate between the last two sets of states if we knew the true generative model and thus, the true difference in the Q-functions, $Q_2(s_2, 1) - Q_2(s_2, 2)$. To get around this, we use the Q-function learned on the original (non-bootstrapped) data to estimate the true Q-function and the bootstrap samples to replicate training data sets.

Algorithm 4 gives the necessary implementation details for constructing the confidence intervals using the bootstrap. Many of the quantities used in Algorithm 4 are used to produce a standard bootstrap confidence interval (this includes the terms $\hat{\mathbb{W}}^{b,m}$, $\hat{\Sigma}_1^{b,m}$ and z_1). In the first set of states, the max operator is not being applied near a point of non-differentiability so standard bootstrap methods apply. The vector z_1 in Algorithm 4 is used by the standard bootstrap procedure. The adjustments to the standard bootstrap are contained in the vectors $\{z_2, z_3, z_4\}$. In the second set of states, the max operator is applied close to or at a point of non-smoothness, and the bootstrap procedure must be modified. Vector z_2 is used to modify the bootstrap for this second set of states. In the third case, the evidence in the data is insufficient to detect the unique optimal action. The vectors z_3 and z_4 along with

Algorithm 3 Calculate Q function and the pretest statistic given estimates $\hat{\beta}_2$, $\hat{\Sigma}_2$

Input: State s_2; Possible treatment sets $\mathcal{A}_2 = \{1, 2\}$; Estimates $\hat{\beta}_2$ and $\hat{\Sigma}_2$; Size of training data set n.

$$\mathcal{T}(s_2) = \frac{n}{(\log(n))} \frac{(\hat{\beta}_2^{\mathsf{T}}(x(s_2, 1) - x(s_2, 2)))^2}{(x(s_2, 1) - x(s_2, 2))^{\mathsf{T}} \hat{\Sigma}_2 (x(s_2, 1) - x(s_2, 2))}.$$

for $a_2 \in \mathcal{A}_2$ **do**

$$\hat{Q}_2(s_2, a_2) = \hat{\beta}_2^{\mathsf{T}} x(s_2, a_2)$$

end for
return $\{\hat{Q}_2(s_2, a_2); a_2 \in \mathcal{A}_2\}, \mathcal{T}(s_2)$.

the maximum in the upper bound (minimum in lower bound) in Algorithm 4 are used to make the appropriate adjustments for this third set of states. If the second and third sets of states were empty ($z_2 = z_3 = z_4 = 0$) then the lower bound $\mathcal{L}^b$, and upper bound $\mathcal{U}^b$ in Algorithm 4 would be equal, and the confidence interval produced by Algorithm 4 would be a standard bootstrap confidence interval. Since the CATIE data set has missing data, the bootstrap samples also have missing data; Algorithm 4 uses Algorithm 1 to impute the missing data in each of the bootstrap samples.

The adaptive confidence interval described here provides a confidence interval with guarantees on coverage probability (see Laber et al. 2010) for a full discussion and justification); in particular, a 95% adaptive confidence interval will cover the true Q-value in about 95% of training sets. The provision of these types of confidence statements are familiar and expected in analyses of clinical trial data. These types of confidence statements are used to decide whether it is worthwhile to conduct future studies of the treatments and state variables. Furthermore, these confidence statements are used to indicate the shortcomings of the investigated treatments (a treatment may need adaption if a particular reward is of interest). The primary drawback of the adaptive confidence interval is that it is complex and somewhat more computationally intensive than other interval estimators appearing in the RL literature.

5 Results

In this section we apply the methods described in Sect. 4 to the CATIE data set. We begin by learning the optimal treatment policy for treating patients with schizophrenia using the CATIE study, and then go on to quantify the evidence for this policy and estimate the uncertainty around the expected outcome for patients who follow this policy. The total reward is the negative of the area under the PANSS score curve over the 18 months of the CATIE study. This is broken up into a reward for the first time interval (the negative of the area under the curve for stage 1) plus the reward for the second time interval (the negative of the area under the curve for stage 2). Recall that a low PANSS score indicates low symptom levels, thus for this application we are interested in minimizing the area under the PANSS curve, or equivalently maximizing the negative of the area under the PANSS curve.

Recall that in the CATIE study, the possible action choices for initial treatment in stage 1 are olanzapine, perphenazine, quetiapine, risperidone and ziprasidone. In the stage 2 tolerability arm, the treatment options are olanzapine, quetiapine, risperidone and ziprasidone. In the stage 2 efficacy arm, we compare the treatment clozapine with one of the set of treatments {olanzapine, quetiapine, or risperidone}. We compare clozapine with the set of

Algorithm 4 Bootstrap algorithm for estimating the bounds for (95%) confidence intervals for stage 1 treatment effects for state s_1, action $a_1 = k$

Input: Training data set $\mathcal{D}$ of size n; B the number of bootstrap iterations; M the number of desired imputations; $\mathcal{S}_2$ is the state space of all observed states in stage 2 of the training set $\mathcal{D}$; a state s_1, action $k \in \mathcal{A}_1$; $\mathcal{A}_2 = \{1, 2\}$ is the set of treatment actions at stage 2.

for $m = 1$ **to** M **do**
  Impute missing data in $\mathcal{D}$ using Algorithm 1, giving completed training set $\mathcal{D}^m$.
  Apply batch fitted Q-iteration to $\mathcal{D}^m$ giving estimated parameters $\hat{\beta}_t^m$ of $\hat{Q}_t^m$ $\forall t$.
end for
Using the average over M imputations of $\mathcal{D}$, calculate estimates $\hat{\beta}_2^{\mathcal{D}}$, $\hat{\Sigma}_2^{\mathcal{D}}$.
for $s_2 \in \mathcal{S}_2$ **do**
  $\left(\{\hat{Q}_2^{\mathcal{D}}(s_2, a_2); a_2 \in \mathcal{A}_2\}, \mathcal{T}^{\mathcal{D}}(s_2)\right) \triangleq$ the output of Algorithm 3 having as input $\mathcal{A}_2 = \{1, 2\}$, $\mathcal{S}_2$, and estimates $\hat{\beta}_2^{\mathcal{D}}$ and $\hat{\Sigma}_2^{\mathcal{D}}$.
end for
for b in $1 : B$ **do**
  Draw bootstrap sample $\mathcal{D}^b$ from $\mathcal{D}$.
  for $m = 1 : M$ **do**
    Impute missing data in $\mathcal{D}^b$ using Algorithm 1, giving completed training set $\mathcal{D}^{b,m}$.
    Estimate $\hat{\beta}_t^{b,m}$ $\hat{\Sigma}_2^{b,m}$ using the m^{th} imputation of the bth bootstrap $\mathcal{D}^{b,m}$.
    for $s_2 \in \mathcal{S}_2$ **do**
      $\left(\{\hat{Q}_2^{b,m}(s_2, a_2); a_2 \in \mathcal{A}_2\}, \mathcal{T}^{b,m}(s_2)\right) \triangleq$ the output of Algorithm 3 having as input $\mathcal{A}_2 = \{1, 2\}$, $\mathcal{S}_2$, and with estimates $\hat{\beta}_2^{b,m}$ and $\hat{\Sigma}_2^{b,m}$.
    end for
    Define, for the ith trajectory in $\mathcal{D}^{b,m}$, the following scalars:
    $\zeta_{1,i} \triangleq \left(\max_{j \in \mathcal{A}_2} \hat{Q}_2^{b,m}(s_{2,i}, j) - \max_{j \in \mathcal{A}_2} \hat{Q}_2^{\mathcal{D}}(s_{2,i}, j)\right)$,
    $\zeta_{2,i} \triangleq \left(\max_{j \in \mathcal{A}_2} \left(\hat{Q}_2^{b,m}(s_{2,i}, j) - \hat{Q}_2^{\mathcal{D}}(s_{2,i}, j)\right)\right)$.
    $\hat{\Sigma}_1^{b,m} = \frac{1}{n} \sum_{i=1}^n x(s_{1,i}, a_{1,i})^{\mathsf{T}} x(s_{1,i}, a_{1,i})$.
    $\hat{\mathbb{W}}^{b,m} = \left(\hat{\Sigma}_1^{b,m}\right)^{-1} \sqrt{n} \sum_{i=1}^n x(s_{1,i}, a_{1,i})^{\mathsf{T}} \left(r_{1,i} + \max_{j \in \mathcal{A}_2} \hat{Q}_2^{\mathcal{D}}(s_{2,i}, j) - \hat{Q}_1^{\mathcal{D}}(s_1, a_{1,i})\right)$.
    Define the following column vectors:
    $z_1 = \frac{1}{n} \sum_{i=1}^n \left[x(s_{1,i}, a_{1,i})^{\mathsf{T}} \cdot \zeta_{1,i} \cdot \mathbb{1}\left(\mathcal{T}^{b,m}(s_2) > 1\right)\right]$.
    $z_2 = \frac{1}{n} \sum_{i=1}^n \left[x(s_{1,i}, a_{1,i})^{\mathsf{T}} \cdot \zeta_{2,i} \cdot \mathbb{1}\left(\mathcal{T}^{b,m}(s_2) \le 1, \mathcal{T}^{\mathcal{D}}(s_2) \le 1\right)\right]$.
    $z_3 = \frac{1}{n} \sum_{i=1}^n \left[x(s_{1,i}, a_{1,i})^{\mathsf{T}} \cdot \zeta_{1,i} \cdot \mathbb{1}\left(\mathcal{T}^{b,m}(s_2) \le 1, \mathcal{T}^{\mathcal{D}}(s_2) > 1\right)\right]$.
    $z_4 = \frac{1}{n} \sum_{i=1}^n \left[x(s_{1,i}, a_{1,i})^{\mathsf{T}} \cdot \zeta_{2,i} \cdot \mathbb{1}\left(\mathcal{T}^{b,m}(s_2) \le 1, \mathcal{T}^{\mathcal{D}}(s_2) > 1\right)\right]$.
    $c = x(s, k)$.
    $\mathcal{U}^{b,m} = c^{\mathsf{T}} \hat{\mathbb{W}}^{b,m} + \sqrt{n}\left[c^{\mathsf{T}} (\hat{\Sigma}_{1,n}^{b,m})^{-1}(z_1 + z_2) + \max\{c^{\mathsf{T}} (\hat{\Sigma}_{1,n}^{b,m})^{-1} z_3, c^{\mathsf{T}} (\hat{\Sigma}_{1,n}^{b,m})^{-1} z_4\}\right]$.
    $\mathcal{L}^{b,m} = c^{\mathsf{T}} \hat{\mathbb{W}}^{b,m} + \sqrt{n}\left[c^{\mathsf{T}} (\hat{\Sigma}_{1,n}^{b,m})^{-1}(z_1 + z_2) + \min\{c^{\mathsf{T}} (\hat{\Sigma}_{1,n}^{b,m})^{-1} z_3, c^{\mathsf{T}} (\hat{\Sigma}_{1,n}^{b,m})^{-1} z_4\}\right]$.
  end for
  $\mathcal{L}^b = \frac{1}{M} \sum_{m=1}^M \mathcal{L}^{b,m}$.
  $\mathcal{U}^b = \frac{1}{M} \sum_{m=1}^M \mathcal{U}^{b,m}$.
end for
Let l be the 2.5^{th} percentile of $\mathcal{L}$ and u the 97.5^{th} percentile of $\mathcal{U}$.
Construct 95% confidence interval for value of action, $a_1 = k$ for state s_1

$$L(s, k) = \hat{Q}_1^{\mathcal{D}}(s_1, k) - l/\sqrt{n},$$

$$U(s, k) = \hat{Q}_1^{\mathcal{D}}(s_1, k) + u/\sqrt{n}.$$

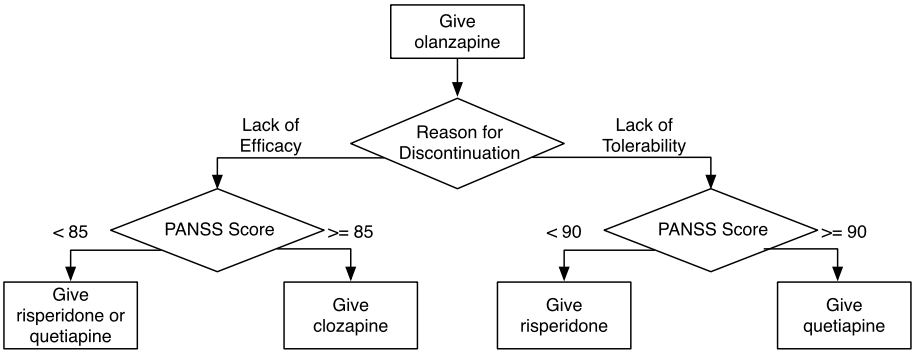

Fig. 3 Optimal treatment policy learned from 25 imputations of the CATIE data, with the total reward defined as the negative area under the PANSS curve for the 18 months of the CATIE study. The state representation is defined in Sect. 3.1 and the Q-function form used is described in Sect. 4.2

treatments for three reasons: (1) the randomization process lends itself to this comparison (50% to clozapine and 16.6% each to olanzapine, quetiapine and risperidone); (2) very few patients enter the efficacy arm leaving us with small amounts of data to learn from; and (3) as previously noted, clozapine is a very different drug from the other three (see Sect. 3), which biologically act similarly.

Figure 3 visually displays the optimal treatment policy learned using 25 imputations of the CATIE data. This learned policy recommends that everyone initially be treated with olanzapine. In the event that an individual chooses to discontinue treatment, the recommended second line of treatment depends on an individual's reason for discontinuing previous treatment and their PANSS score at that time.

Figure 4 presents the bootstrap voting plots for stage 1 of the CATIE study. The bootstrap voting plots for the stage 2 efficacy and tolerability arms are shown in Fig. 5. In each of the bootstrap voting plots, the vertical location of the diamonds indicates the estimated value of the optimal action. The size of the diamonds shows the number of people in the training set within the range of PANSS scores indicated on the horizontal axis. Figures 4 and 5(b) both convey information about the relative evidence for five different actions, along with the estimated value for the optimal action. We can readily see from Fig. 4 that there is strong evidence for olanzapine as the optimal initial treatment. In Figs. 5(a) and (b) it is clearly seen that the data provides evidence suggesting that the optimal treatment choice should change as a function of PANSS score. In Figs. 4 and 5(b), a portion of the bars is unshaded. This area represents the estimate for $p_{\emptyset}$, the proportion of discordant results. In Fig. 4 the estimate for $p_{\emptyset}$ increases with PANSS (note also that sample size decreases as PANSS increases in both figures) while in Fig. 5(b) the estimated proportion of discordant results remains fairly constant. This illustrates that it is not only training set size, but also small treatment effects and variable data, that contribute to the lack of evidence for a unique optimal treatment. In Fig. 4(a) there is no unshaded area; this is because when there are only two action choices discordant results are not possible: one of the actions must be optimal.

Figures 6 and 7 visually present the results of applying the adaptive confidence interval methodology to the CATIE data. The circles represent the point estimate of the value for each action and PANSS score, while the confidence interval is indicated by a vertical line. From Fig. 6 it is immediately clear that while the point estimate for value function of olanzapine is always lowest (recall a low PANSS score is good), its confidence intervals always overlap with the confidence intervals for at least one other treatment. This indicates that our

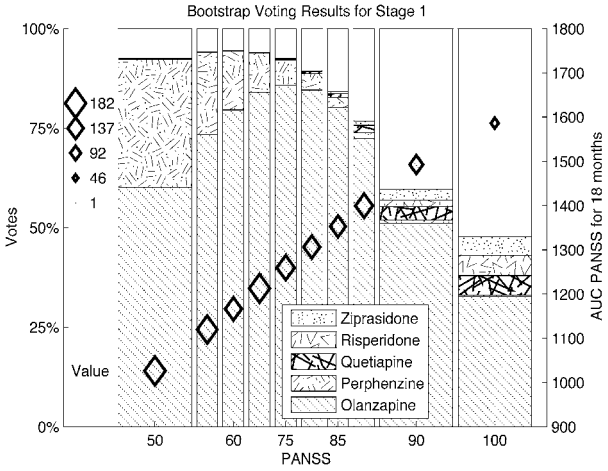

Fig. 4 Bootstrap voting results for stage 1 of CATIE study. The *vertical location of the diamonds* indicate the point estimates for the value function associated with different stage 1 PANSS scores. The *vertical axis on the right hand* side gives the axis for this value function. The *size of the diamonds* represents the number of people in the corresponding PANSS bin in the CATIE study. The *shaded bar* represents the evidence in the data for each of the action choices as labeled in the *legend*. The *unshaded portion of the graph* is the estimate for $\hat{p}_{\emptyset}$, the proportion of discordant trials. Recall a low PANSS score indicates low symptoms of schizophrenia, thus a lower score is better

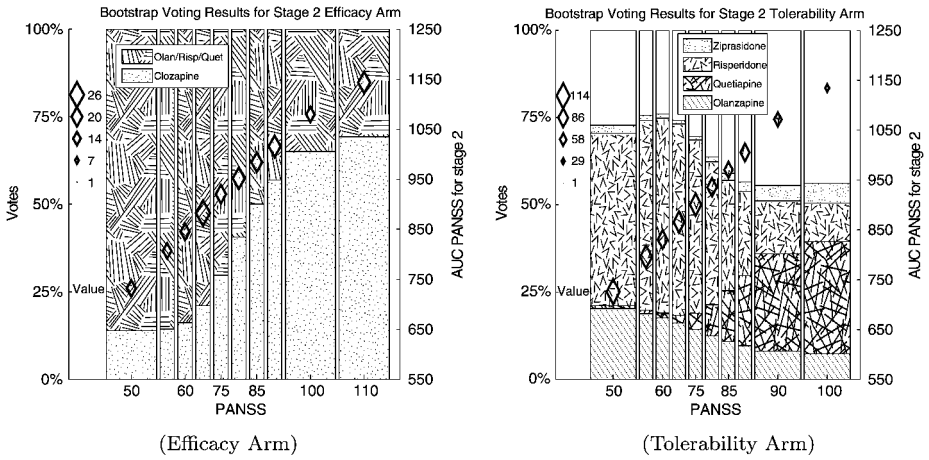

Fig. 5 Bootstrap voting results for stage 2 of CATIE study. Plots for the efficacy and tolerability are in figures (**a** Efficacy Arm) and (**b** Tolerance Arm) respectively. The *vertical location of the diamonds* in these plots represent the point estimates for the value function with different stage 2 PANSS scores. The *vertical axis on the right hand side* gives the axis for the value function. The *size of the diamonds* represents the number of people in the corresponding PANSS bin in the CATIE study. The *shaded regions* represent evidence for each of the action choices labeled in the legends, $\hat{p}_k^{win}$. The *unshaded portion of the graph* is the estimate for $\hat{p}_{\emptyset}$, the proportion of discordant trials. Recall a low PANSS score indicates low schizophrenia symptoms, thus a lower score is better

uncertainty about the value of choosing olanzapine as the initial treatment is high, and we cannot with high confidence recommend olanzapine as the sole optimal initial treatment. At the same time, if we focus on the lowest PANSS score, the confidence interval for olanzap-

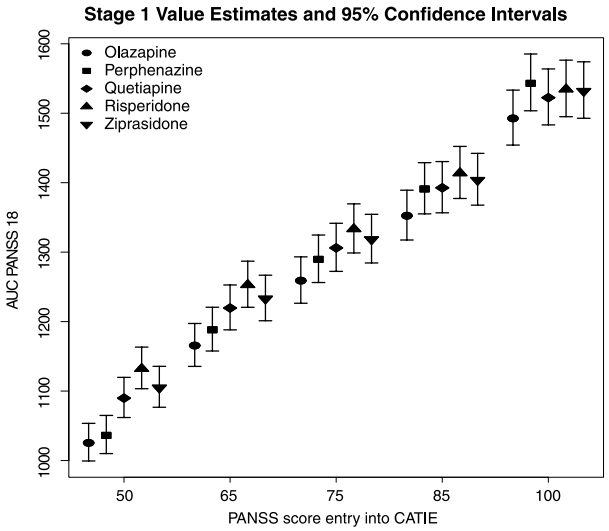

Fig. 6 Estimates and 95% confidence intervals for stage 1 state-action value-function. The *circle* represents the point estimate for the value of each action given the PANSS score indicated on the *horizontal axis*. The different stage 1 treatments are represented by the colors indicated in the *legend*

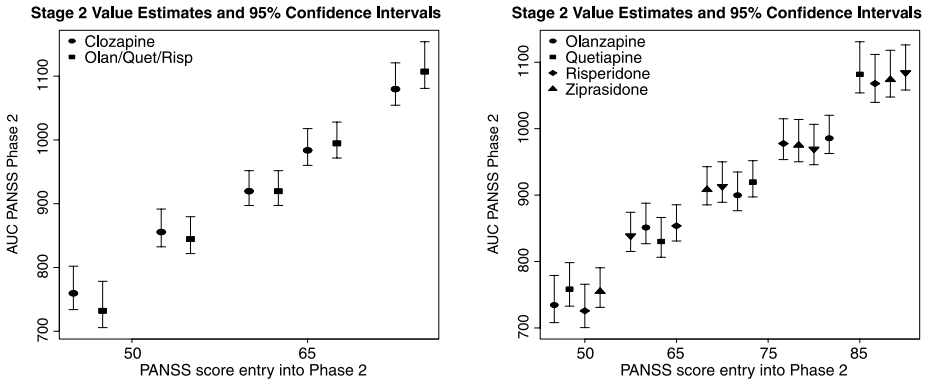

Fig. 7 Estimates and 95% confidence intervals for stage 2 state-action Q-function. The circle represents the point estimates of the state-action value function for each action given the PANSS score indicated on the *horizontal axis*. The various stage 2 treatments are represented by the colors indicated in the *legend*

ine does *not* overlap with risperidone, quetiapine or ziprasidone. Thus, these three drugs can be excluded as possible candidates for the optimal action for individuals with low PANSS scores. Even more uncertainty exists for the value function estimates for the second stage of CATIE. In the efficacy arm, we can see that the point estimate for clozapine is above the other treatment group at low PANSS values (indicating it is less effective), but it is below the others at high PANSS values. However, the confidence intervals around these estimates overlap, indicating a lack of confidence in either treatment. Similar results hold in the tolerance arm where the point estimate for the value of risperidone is best for low PANSS scores and for high PANSS scores the point estimate for the value of quetiapine is better. Again, the overlapping confidence intervals lead us to conclude that we do not have high confidence that there is one unique best treatment action.

6 Discussion

This paper highlights the role that reinforcement learning can play in the optimization of treatment policies for informing clinical decision making. We have presented methods for tackling some of the important issues that often arise when applying reinforcement learning to data collected in a SMART study. Finally, we showed the results of applying these methods to real data collected from a SMART study for patients with schizophrenia. We now conclude the paper by discussing some general applications of these methods, outlining some of limitations of these methods, and detailing some interesting open questions.

6.1 Relevance of RL in sequential treatment design

The use of RL methods to automatically learn good treatment policies for patients with chronic illnesses is an emerging field, with great opportunities for new research directions. A number of SMART studies are currently underway in the fields of autism (C. Kasari, personal communication), alcoholism (D.W. Oslin, pers. comm.; J.R. Mackay, pers. comm.), drug abuse (H. Jones, pers. comm.), and attention deficit disorder (W.E. Pelham, pers. comm.). RL provides the basic principles and mathematical foundations for learning evidence-based, individualized, treatment policies from clinical data.

While this foundation is a good start, using RL methods to optimize treatment policies is not as simple as applying off-the-shelf RL methods. Though the planning horizon is very short (only 2 stages in our particular application) and the exploration policy is fixed, upon closer inspection, a number of previously unexplored challenges arise when tackling problems in this domain. In particular, our case study highlights issues such as pervasive missing data, the need to handle a high-dimensional and variable state space and the need to communicate the evidence for a recommended action and estimate confidence in the actions selected by the learned policy. The methods we present in Sect. 4, and apply to the CATIE data set to tackle these problems, are of particular interest in this domain. They are likely useful in a wide range of other domains and may in fact open new avenues of research in the RL field.

6.2 Bootstrap voting plots and adaptive confidence intervals

The bootstrap voting procedure allows us to clearly illustrate our best estimate of the probability of selecting a particular action in an analysis on a future training set. While the bootstrap voting procedure has this meaningful probabilistic interpretation, since we are only giving a point estimate of this quantity, the results cannot be interpreted as a measure of confidence. The usefulness of the procedure is in conveying the evidence in our current data set for the action choices made at different states, and in particular, clearly showing how the evidence changes with state. One direction for future work would be to produce confidence intervals for the bootstrap voting estimator and devise a means of augmenting the bootstrap voting plots with this confidence information. Note that the adaptive confidence intervals presented in Sect. 5 show our confidence about the predicted value of each action. In order to compare two actions, we would construct a confidence interval for the difference in action value and determine whether that confidence interval included zero. To compare all of the actions in stage 1 of CATIE would thus require $\binom{5}{2} = 10$ confidence intervals at each state value, which becomes unwieldy to present. This is one of the reasons we have presented both the bootstrap voting plots and the adaptive confidence intervals results. While the bootstrap voting method does not convey confidence, it does convey evidence for action selection, and it does so very concisely for any number of actions.

The adaptive confidence interval method presented here assumes that the linear approximation provides a high quality approximation to the optimal Q-function. The adaptive confidence interval method has not yet been generalized for use with a non-linear approximation to the Q-function such as, for example, trees or a nearest neighbor estimators. Confidence intervals are particularly useful in evidence-based medical decision-making, but constructing such intervals could be very beneficial in other areas in which the same training set must be used to both learn a policy and then evaluate the learned policy. Interesting extensions of this work would involve extending these results to a richer class of function approximators.

6.3 Problems of missing data

As we have pointed out in this paper, data from clinical trials are often characterized by missing data problems. The patterns and amount of missingness can vary greatly by disease. For example it is typical for mental illness trials to have higher dropout rates than cancer clinical trials. While it is tempting to simply ignore trajectories with missing information, this is not a good strategy. Removing individuals with missing data from a training set increases the variance of estimators by reducing the training set size and can add bias to the estimates of the action effects on the Q-function.

The reinforcement learning community has usually turned to POMDPs to tackle problems of missing data. While the mathematical framework is appropriate for many domains, the usefulness of POMDPs in practical applications with many variables remains problematic due to the complexity of the inference problem. In this paper, we highlight how methods from the statistical literature can be leveraged to overcome the missing data problem and produce fully observable (imputed) trajectories. Such methods may be useful in a wider range of RL applications where missing data occurs.

The multiple imputation framework itself is very general and can be implemented with very few modeling assumptions. Although, as in the CATIE example, it is often useful to make appropriate modeling assumptions to ensure tractability. These assumptions also provide an opportunity for including existing domain knowledge, about the causes of the missingness and the structure of the data. As with any assumptions made in the modeling process, the validity of the assumptions should be tested and where they cannot be tested, sensitivity analyses should be preformed. There is a large literature on testing the validity of the assumptions made in multiple imputation, as well as a variety of sensitivity analyses (Little and Rubin 1987; Robins et al. 1999; Scharfstein et al. 1999; Gelman et al. 2005; Carpenter et al. 2007; NAP 2010).

6.4 Designing the reward function

As briefly mentioned in Sect. 3, one of the important open problems in applying RL to learning treatment policies is the definition of the reward. In our particular example, we chose to use the negative of the area under the curve of the PANSS score over the 18 months of the CATIE study. A composite reward that accounts for symptoms over a long period of time is important in the treatment of schizophrenia. As with many chronic diseases, symptoms vary over time, and reducing a patient's symptoms at a single point in time does not ensure that their symptoms will remain low. The AUC is one example of a composite score, but other choices for the reward function are possible and could also prove to be interesting. Choosing the reward function is a non-trivial task and should be done in collaboration with clinicians to ensure that policies learned are both interpretable and able to be effectively implemented in practice.

In this paper we focus on learning a policy that minimizes symptoms as measured by the PANSS score. In most clinical settings, treatment decisions are based on a rich interplay of factors, including patient response to previous treatment, the development of new symptoms and illnesses, and side-effects among others. A limitation of applying the traditional RL framework is that it focuses on a single (scalar) reward. The medical setting stresses the importance of the development of RL methods that can optimize performance considering a set of (possibly competing) rewards. A few researchers have explored special cases of this problem (Bagnell et al. 2001; Shelton 2001; Doshi et al. 2008; Lizotte et al. 2010) but more development is required.

6.5 Informing data collection

Working with batch data has the obvious advantage that the exploration issue is addressed *a priori*. The traditional randomized clinical trial collects data under a pure exploration policy; however, not all trials follow this strategy. There is a growing interest in the medical community in conducting *Bayesian adaptive clinical trials* (Berry 2006; Thall and Wathen 2007; Biswas et al. 2009), in which some aspects of the trial design (e.g. randomization probabilities, trial population) are changed adaptively throughout the trial, based on the information collected. This approach can be used to make more efficient use of data and could help ensure that patients receive the best possible treatment, even during the early stages of treatment testing. In the reinforcement learning literature, Bayesian approaches have also been explored for similar reasons (Dearden et al. 1999; Strens 2000; Wang et al. 2005; Engel et al. 2005); other methods targeting efficient exploration have also been explored (Kakade et al. 2003; Strehl et al. 2006; Brunskill et al. 2008). The medical domain is likely to be a fruitful application area for these techniques.

An important concern with training data acquired through clinical trials is the fact that all actions are potentially equally good. Recall that the ethical issues mentioned in Sect. 3.3 prohibit the exploration of treatments known to be less effective than existing treatments. This implies that the difference between Q-values for different actions is likely to be small. Furthermore, we need to estimate Q-values from relatively few trajectories. There is no easy solution to this challenge, rather we believe it is particularly important to *know what we don't know*. Clinical trials are expensive to run, yet as new treatments emerge, new trials are run to compare new treatments to current ones. Ruling out treatments as non-optimal is informative in the medical setting. For example, in the CATIE study, while we do no have enough evidence to recommend one optimal treatment, the adaptive confidence intervals in Fig. 6 indicate that quetiapine, risperidone and ziprasidone are all non-optimal for low PANSS scores. Thus, these three treatments can be ruled out as a best initial treatment for patients with low PANSS scores. This information can be very important for informing the design of future clinical trials. Recently, researchers have begun to develop methods for estimating sets of actions that can be classified as near-optimal in the RL setting (Fard and Pineau 2009). Medical applications open the door to these new avenues of research focused on estimating sets of nearly optimal treatments rather than simply focusing on learning a single optimal treatment at each stage.

Acknowledgements We acknowledge support from the National Institutes of Health (NIH) grants R01 MH080015 and P50 DA10075, the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Institutes of Health Research (CIHR).

References

Adams, C. E. (2002). Schizophrenia trials: past, present and future. *Epidemiologia E Psichiatria Sociale*, *11*(13), 144–151.

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, *68*(2), 399–405.

Bagnell, A., Ng, A., & Schneider, J. (2001). *Solving uncertain Markov decision problems* (Tech. Rep. CMU-RI-TR-01-25). Robotics Institute, Carnegie Mellon University.

Berry, D. A. (2006). A guide to drug discovery: Bayesian clinical trials. *Nature Reviews. Drug Discovery*, *5*, 27–36.

Biswas, S., Liu, D. D., Lee, J. J., & Berry, D. A. (2009). Bayesian clinical trials at the University of Texas M. D. Anderson cancer center. *Clinical Trials*, *6*, 205–216.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Brunskill, E., Leffler, B. R., Li, L., Littman, M., & Roy, N. (2008). A continuous-state offset-dynamics reinforcement learner. In D. A. McAllester & P. Myllymäki (Eds.), *Proceedings of 24th conference on uncertainty in artificial intelligence (UAI 2008)* (pp. 53–61).

Carpenter, J. R. , Kenward, M. G., & White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research*, *16*(3), 259–275.

Dawson, R., & Lavori, P. W. (2004). Placebo-free designs for evaluating new mental health treatments: the use of adaptive strategies. *Statistics in Medicine*, *23*, 3249–3262.

Dearden, R., Friedman, N., & Andre, D. (1999). Model based Bayesian exploration. In B. Kathryn, & H. P. Laskey (Eds.), *Proceedings of 5th conference on uncertainty in artificial intelligence (UAI 1999)* (pp. 150–159). San Mateo: Morgan Kaufmann.

Diggle, P., Heagerty, P., Liang, K. Y., & Zeger, S. (2002). *Analysis of longitudinal data*. Oxford: Oxford University Press.

Doshi, F., Pineau, J., & Roy, N. (2008). Reinforcement learning with limited reinforcement: using Bayes risk for active learning in POMDPs. In A. McCallum & S. Roweis (Eds.), *Proceedings of the 25th annual international conference on machine learning (ICML 2008)* (pp. 256–263). New York: Omnipress.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Engel, Y., Mannor, S., & Meir, R. (2005). Reinforcement learning with Gaussian processes. In L. D. Raedt & S. Wrobel (Eds.), *Proceedings of the 22nd international conference on machine learning (ICML 2005)* (pp. 201–208). New York: ACM. 10.1145/1102351.1102377.

Ernst, D., Geurts, P., & Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, *6*, 503–556.

Ernst, D., Stan, G. B., Goncalves, J., & Wehenkel, L. (2006). Clinical data based optimal STI strategies for HIV: a reinforcement learning approach. In *Proceedings of the machine learning conference of Belgium and The Netherlands (Benelearn)* (pp. 65–72).

Fard, M. M., Pineau, J. (2009). MDPs with non-deterministic policies. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems* (pp. 1065–1072). Cambridge: MIT Press.

Gelman, A., Carlin, J. B., Stern, H., & Rubin, D. B. (1995). *Bayesian Data Analysis*. New York: Chapman & Hall.

Gelman, A., Mechelen, I. V., Verbeke, G., Heitjan, D. F., & Meulders, M. (2005). Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics*, *61*, 74–85.

Guez, A., Vincent, R., Avoli, M., & Pineau, J. (2008). Adaptive treatment of epilepsy via batch-mode reinforcement learning. In *Proceedings of the innovative applications of artificial intelligence (IAAI)*.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *Elements of statistical learning*. Berlin: Springer.

Irodova, M., & Sloan, R. H. (2005). Reinforcement learning and function approximation. In *Proceeding of the twentieth national conference on artificial intelligence (AAAI)* (p. 2005). American Association for Artificial Intelligence, Menlo Park.

Kaelbling, L. P., Littman, M. L., & Moore, A. (1996). Reinforcement learning: a survey. *The Journal of Artificial Intelligence Research*, *4*, 237–385.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*, 99–134.

Kakade, S., Kearns, J., & Langford, J. (2003). Exploration in metric state spaces. In *Proceedings of the 20th Annual International Conference on Machine Learning (ICML 2003)*.

Kay, S. R., Flazbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, *13*(2), 261–276.

Laber, E. B., Qian, M., & Murphy, S. A. (2010). *Statistical inference in dynamic treatment regimes* (Tech. Rep. 506). Dept. of Statistics, University of Michigan

Lagoudakis, M. G., & Parr, R. (2003). Least-squares policy iteration. *Journal of Machine Learning Research*, *4*, 1107–1149.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Lizotte, D. J., Laber, E., & Murphy, S. A. (2009) *Assessing confidence in policies learned from sequential randomized trials* (Tech. Rep. 481). Department of Statistics, University of Michigan.

Lizotte, D., Bowling, M., & Murphy, S. (2010). Efficient reinforcement learning with multiple reward functions for randomized controlled trial analysis. In *Proceedings of the twenty-seventh international conference on machine learning (ICML 2010)*. (pp. 695–702). New York: Omnipress.

Mannor, S., Simester, D., Sun, P., & Tsitsiklis, J. (2007) Biases and variance in value function estimates. *Management Science* 53(1).

Monahan, G. (1982). A survey of partially observable Markov decision processes. *Management Science*, *28*, 1–16.

Murphy, S. M. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B*, *65*(2), 331–366.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, *24*, 1455–1481.

Murphy, S. A., Oslin, D., & Rush, A. J. (2007). Methodological challenges in constructing effective treatment sequences for chronic disorders. *Neuropsychopharmacology*, *32*(2), 257–262.

NAP (2010). *The prevention and treatment of missing data in clinical trials*. The National Academies Press, Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral, Social Sciences and Education.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models*. New York: McGraw-Hill.

Parr, R., Li, L., Taylor, G., Painter-Wakefield, C., & Littman, M. (2008). An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In A. McCallum, & S. Roweis (Eds.), *Proceedings of the 25th annual international conference on machine learning* (pp. 752–759). New York: Omnipress.

Pineau, J., Bellemare, M. G., Rush, A. J., Ghizaru, A., & Murphy, S. A. (2007). Constructing evidence-based treatment strategies using methods from computer science. *Drug and Alcohol Dependence* S52–S60.

Robins, J. M., Rotnitzky, A., & Scharfstein, D. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In M. E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology: the environment and clinical trials* (pp. 1–92). Berlin: Springer.

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, *91*, 473–489.

Rush, A. J., Fava, M., Wisniewski, S. R., & Lavori, P. W. (2004). Sequenced treatment alternatives to relieve depression (STAR*D): rational and design. *Controlled Clinical Trials*, *25*(1), 119–142.

Schafer, J. L. (1997). *Imputation of missing covariates under a multivariate linear mixed model* (Tech. rep.). Dept. of Statistics, The Pennsylvania State University.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, *8*(1), 3–15.

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed models with missing values. *Journal of Computational and Graphical Statistics*, *11*, 421–442.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, *94*(448), 1096–1120.

Shao, J. (1994). Bootstrap sample size in nonregular cases. *Proceedings of the American Mathematical Society*, *122*(4), 1251–1262.

Shelton, C. R. (2001). Balancing multiple sources of reward in reinforcement learning. In *Advances in neural information processing systems (NIPS 2000)* (pp. 1082–1088).

Smallwood, R. D., & Sondik, E. J. (1973). The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, *21*, 1070–1088.

Strehl, A. L., & Littman, M. L. (2004). An empirical evaluation of interval Estimation for Markov decision processes. In *ICTAI* (pp. 128–135). Los Alamitos: IEEE Computer Society.

Strehl, A. L., & Littman, M. L. (2005). A theoretical analysis of model-based interval Estimation. In L. D. Raedt & S. Wrobel (Eds.), *Proceedings of the 22nd international conference on Machine learning (ICML 2005)* (pp. 856–863). New York: ACM. 10.1145/1102351.1102459.

Strehl, A., Li, L., Wiewiora, E., Langford, J., & Littman, M. (2006). PAC model-free reinforcement learning. In W. W. Cohen & A. Moore (Eds.), *Proceedings of the 23rd annual international conference on machine learning (ICML 2006)* (pp. 881–888).

Strens, M. J. A. (2000). A Bayesian framework for reinforcement learning. In P. Langley (Ed.), *Proceedings of the seventeenth international conference on machine learning (ICML 2000)* (p. 943–950). San Francisco: Morgan Kaufmann.

Stroup, T. S., McEvoy, J. P., Swartz, M. S., Byerly, M. J., Glick, I. D., Canive, J. M., McGee, M., Simpson, G. M., Stevens, M. D., & Lieberman, J. A. (2003). The National Institute of Mental Health clinical antipschotic trials of intervention effectiveness (CATIE) project: schizophrenia trial design and protocol development. *Schizophrenia Bulletin*, *29*(1), 15–31.

Sutton, R. S., & Barto, A. G. (1998). Off-policy bootstrapping. In *Reinforcement learning: an introduction* Cambridge: MIT Press.

Swartz, M. S., Perkins, D. O., Stroup, T. S., McEvoy, J. P., Nieri, J. M., & Haal, D. D. (2003). Assessing clinical and functional outcomes in the clinical antipsychotic of intervention effectiveness (CATIE) schizophrenia trial. *Schizophrenia Bulletin*, *29*(1), 33–43.

Tetreault, J., Bohus, D., & Litman, D. (2007). Estimating the reliability of MDP policies: a confidence interval approach. In *Proceedings of the human language technology conference* (pp. 276–283).

Thall, P., & Wathen, J. (2007). Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cancer*, *43*(5), 859–866.

Thall, P. F., & Wathan, J. K. (2000). Covariate-adjusted adaptive randomization in a sarcoma trial with multistate treatments. *Statistics in Medicine*, *19*, 1011–1028.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064.

Wang, T., Lizotte, D., Bowling, M., & Schuurmans, D. (2005). Bayesian sparse sampling for on-line reward optimization. In L. D. Raedt & S. Wrobel (Eds.), *Proceedings of the 22nd international conference on machine learning (ICML 2005)* (pp. 956–963). New York: ACM. 10.1145/1102351.1102472.

Zhao, Y., Kosorok, M. R., & Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in Medicine*, *28*, 3294–3315.