Visualizing Clinical Significance with Prediction and Tolerance Regions

Maria Jahja MJAHJA@NCSU.EDU
Department of Statistics
North Carolina State University
Raleigh, NC 27695

Daniel J. Lizotte DLIZOTTE@UWO.CA
Department of Computer Science
Department of Epidemiology & Biostatistics
University of Western Ontario
London, ON N6A 3K7

Abstract

The goal of this work is to better convey the evidence for or against clinically significant differences in patient outcomes induced by different treatment policies. In pursuit of this goal, we present a framework for computing and presenting prediction regions and tolerance regions for the outcomes of a treatment policy operating within a multi-objective Markov decision process (MOMDP). Our framework draws on two bodies of existing work, one in computer science for learning in MOMDPs, and one in statistics for uncertainty quantification. We review the relevant methods from each body of work, present our framework, and illustrate its use using data from the Clinical Antipsychotic Trials of Intervention Effectiveness (Schizophrenia). Finally, we discuss potential future directions of this work for supporting sequential decision-making.

1. Introduction

In its broadest sense, *clinical significance* describes the degree to which a treatment will make a meaningful impact on a patient's health outcomes. In contrast, *statistical significance* measures the degree to which an observed average difference in outcome, estimated using a group of patients, could have arisen by chance alone (Porta, 2014). These two are not necessarily aligned; one can observe a statistically significant average difference—even if that difference is very small—so long as the dataset under consideration is big enough. As dataset sizes have grown, both from a push toward large pragmatic clinical trials and from the increased availability of large observational datasets, there is a much greater opportunity than before to observe statistical significance without clinical significance.

Consider the simplified analysis of CATIE, the Clinical Antipsychotic Trials of Intervention Effectiveness, shown in Table 1. The table shows regression analyses of symptom (Positive and Negative Syndrome Score—PANSS) and side-effect (Body Mass Index—BMI) scores at the end of the study, adjusting for baseline covariates, and comparing two first-line treatments, ziprasidone and olanzapine ($n = 364$). In the analyses, the impact of ziprasidone versus olanzapine is statistically significant; on average, patients on ziprasidone had a

Table 1: Olanzapine versus ziprasidone as first-line treatment

PANSS Model	Estimate	p-value	**BMI Model**	Estimate	p-value
(Intercept)	19.07516	< 0.0001	(Intercept)	2.2243	0.0024
Baseline.PANSS	0.66216	< 0.0001	Baseline.BMI	0.9663	< 0.0001
Ziprasidone	4.46974	0.0127	Ziprasidone	-0.9598	0.0054

higher (worse) PANSS score, and a lower (better) BMI than patients on olanzapine, after adjusting for baseline levels. However, as Figure 1 illustrates, the actual observed outcomes are highly variable, and there is a great deal of overlap between the two treatment groups.

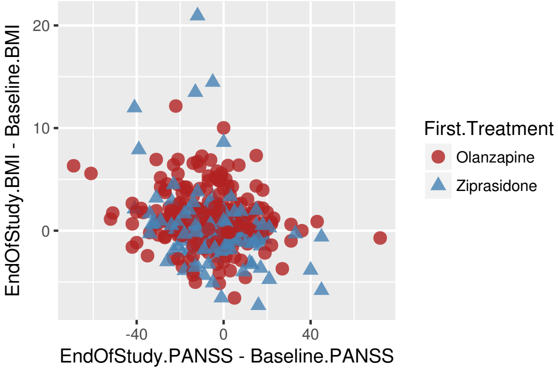

Figure 1: Observed changes in outcomes

Our work aims to convey visually what outcomes are likely for a given individual under different candidate treatment policies. Our visualization can take into account multiple outcomes of interest simultaneously, and applies to sequential decision-making problems formalized as a Markov Decision Processes (MDP) (Bertsekas, 2007). (Although the CATIE study involved sequences of treatments, Figure 1 considers only the first-line treatment and averages over future treatments rather than optimizing.) MDPs are useful conceptual tools for reasoning about sequential decision-making under uncertainty. Much of the computer science research on planning and learning in MDPs has focused on constructing an autonomous agent that acts in a given environment over an extended period of time, choosing actions according to a particular policy in order to achieve a high expected sum of rewards (Sutton and Barto, 1998). Other research, particularly in statistics, uses MDPs to frame the development of evidence-based decision support for sequential decision-making problems (Laber et al., 2014). In the field of statistics, policies are often called *Dynamic Treatment Regimes* (DTRs), and there is a substantial literature studying their development and application in the field of health care decision-making (Orellana et al., 2010; Shortreed et al., 2011). As in computer science, much of the literature is devoted to the estimation of regimes that optimize the expected sum of rewards, as well as to uncertainty quantification for the parameters and the *average* performance of such regimes (Laber et al., 2014; Lizotte and Tahmasebi, 2017).

Most DTR literature focuses on the use of batch data to understand how treatments can be selected with the goal of achieving long-term success for a population of patients. Thus, the deployment of DTRs in statistics was always assumed to be "multi-agent" in the sense that the DTR would be used to aid the decisions of *many different patient-clinician pairs* where each patient would experience only one "episode" of the regime. Thus, there is a fundamental disconnect between the *average* performance of a regime ("How will a population respond?") and the *individual* performance of a regime ("How will an individual respond?"). In a computer science framework with a single agent over a long horizon, this difference matters much less; in decision support, however, recognizing the variability in performance achieved over individual episodes and communicating that variability to human decision-makers is crucial. The difference, as discussed previously, can be viewed as one of statistical versus clinical significance—with enough data, we may become extremely confident that one action affords higher *expected* return than another in a given state. However, if the variance of the returns is large, the question of which action will perform better in a particular episode may nearly come down to a fair coin flip.

Our goal is to capture and convey information about the *distribution* of returns, rather than only the mean return, of a learned policy. We also wish to accommodate multiple reward signals; rather than formulate an MDP using a single, one-size-fits-all reward, we use a *Multi-Objective Markov Decision Process* (MOMDP) to enable the end user to consider several rewards that may be important (e.g. symptom relief, side effects, cost) before making their next decision. To accomplish this goal, we extend and combine ideas from recent work on uncertainty quantification in MDPs and on Q-learning for MOMDPs. Lizotte and Tahmasebi (2017) present methods for constructing *prediction intervals* and *tolerance intervals* for the returns of a policy. Their method conveys, given a state and action, what returns are likely to be observed, but is restricted to the single-reward setting. Lizotte and Laber (2016) describe a methodological and computational framework for computing optimal policies of MOMDPs under different solution concepts using linear function approximation over a finite time horizon. Their approach, an extension of Q-learning, provides point estimates for the mean vector-valued returns achievable from a given state, but does not give information about the distribution of vector-valued returns.

Our main technical contribution is a framework for computing tolerance regions in the multiple-reward setting by augmenting both the tolerance interval algorithm from Lizotte and Tahmasebi (2017) and the policy learning algorithm from Lizotte and Laber (2016). The output of our algorithm is the region of the space of returns that is most likely to contain a return achieved by following a non-dominated policy. We present a framework rather than a particular algorithm, because different applications are likely to warrant different components for the Q-function approximation, solution concept, and region construction. We give a clinical example to demonstrate how the framework functions, but we also identify where components could be interchanged to suit different tasks.

An important secondary goal of our work is simply to illustrate how the methodology of prediction and tolerance regions can be used to explore clinical significance and aid decision-making. These methods can be directly applied in simpler settings, such as two-arm randomized clinical trials or observational studies, to summarize the experiences of the patients without requiring some of the more complicated methodology we describe here.

2. Cohort and Outcomes

Our work was motivated by our experience in analyzing large pragmatic clinical trials with multiple stages of randomization and multiple outcomes. We illustrate the output of non-deterministic fitted-Q using data from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) study. The CATIE study was designed to compare sequences of antipsychotic drug treatments for the care of schizophrenia patients. It was designed as a pragmatic trial that was intended to mimic clinical practice as closely as possible; hence, the inclusion criteria were made as broad as was feasible, and the recruitment was large ($n = 1460$). The full study design is quite complex (Stroup and al, 2003; Swartz et al., 2003); we use a simplified subset of the CATIE data in order to more clearly illustrate the proposed methodology. Briefly, participants were randomized to one of a collection of antipsychotics on entry to the study, and were monitored over time. If, during the study, the participant and their physician decided that the current treatment was not acceptable, either due to a lack of efficacy (i.e. symptom reduction) or tolerability (i.e. side effects), they were randomized to a different treatment and were monitored until the end of the study.[1] A more complete description of the protocol is provided in Appendix A, but for a full description we refer the reader to the design published by Stroup and al (2003). To simplify our analysis, we focus only on ($n = 356$) patients who were randomized to olanzapine or ziprasidone at the first phase of the study. These two atypical antipsychotics offer a tradeoff between symptoms and side-effects: as noted in the simplified analysis above, in our dataset olanzapine on average appears to be superior in terms of symptom reduction but inferior in terms of weight gain.

In previous work, the goal of analyzing CATIE has been to develop a two phase policy that consists of a rule for choosing the intial treatment and then a rule for choosing a follow-up treatment if desired. Such work has examined the problem of finding optimal policies in terms of expected outcome over patients, both in the single-outcome (Shortreed et al., 2011) and the multiple-outcome (Lizotte and Laber, 2016) settings. We will address the multiple outcome setting, focussing on PANSS, which measures schizophrenic symptoms, and BMI, which attempts to measure whether a person is of a healthy weight. Weight gain is one of the most significant side-effects of antipsychotic medications (Leucht et al., 2013). Rather than producing a treatment policy, our goal is to provide a decision aid to help guide the choice of first-line treatment in a way that acknowledges the potential effects of future treatment choices.

3. Background

In the following, we review background for Multi-Objective Markov Decision Processes and Q-learning in that context. We then review prediction regions and tolerance regions.

3.1 Multi-Objective Markov Decision Processes and Q-learning

This section follows the development in previous work by Lizotte and Laber (2016). Clinical decision-making is often driven by multiple competing objectives; for example, a medical decision will be based not only on the effectiveness of a treatment, but also on its potential side-effects, cost, and other considerations. Because the relative importance of these objectives varies from individual to individual, the quality of a policy may not be well captured by a universal single scalar “reward” or “value.” Multi-Objective Markov Decision Processes (MOMDPs) accommodate this by allowing vector-valued rewards Lizotte and Laber (2016); Roijers et al. (2013) and using an application-dependent *solution concept* to define the performance of a policy. A solution concept is essentially a partial order on policies; the

[1] These represent complete cases, which simplifies our presentation in the paper. We conducted a multiple-imputation-based version of our simplified analyses in Table 1 and found minimal differences in point estimates and significance levels as compared with the complete case analysis.

set of policies that are maximal according to the partial order are considered "optimal" and are indistinguishable under that solution concept – for example, using Pareto optimality as the solution concept leads to an equivalence class of all policies that lead to a value on the Pareto frontier. A collection of policies may be represented by a *non-deterministic policy* (NDP) Milani Fard and Pineau (2011). Given an MDP with *state space* $\mathcal{S}$ and an *action set* $\mathcal{A}$, an NDP Π is a map from the state space to the set $2^{\mathcal{A}} \backslash \{\varnothing\}$. Like previous work Lizotte and Laber (2016); Roijers et al. (2013), we focus on the setting where the definition of an MDP is augmented by assuming a D-dimensional reward vector $\mathbf{R}(s_t, a_t)$ is observed at each time step. We define a MOMDP with finite time horizon T as a tuple of state spaces $\mathcal{S}_t$, action spaces $\mathcal{A}_t$, state transition functions $P_t : \mathcal{S}_t \times \mathcal{A}_t \to \mathbb{P}(\mathcal{S}_{t+1})$ where $\mathbb{P}(\mathcal{S}_{t+1})$ is the space of probability measures on $\mathcal{S}_{t+1}$, and reward functions $\mathbf{R}_t : \mathcal{S}_t \times A_t \to \mathbb{R}^D$ for $t \in \{1, ..., T\}$. In keeping with the Markov assumption, both $\mathbf{R}_t$ and P_t depend only on the current state and action. We assume finite action sets, but we do *not* assume that state spaces are finite. The *value* of a policy π is given by $\mathbf{V}^{\pi}(s) = \mathbb{E}^{\pi}[\sum_{t=1}^{T} \mathbf{R}^t(s_t, a_t) | s_1 = s]$, the expected sum of (vector-valued) rewards we achieve by following π.

Consider a batch of n trajectories $s_1^i, a_1^i, r_{1[1]}^i, ..., r_{1[D]}^i, , ..., s_T^i, a_T^i, r_{T[1]}^i, ..., r_{T[D]}^i$ for $i = 1, ..., n$. At time T, (the final time point) we define the approximate Q-function for reward dimension d as the least squares fit

$$\hat{Q}_{T[d]}(s_T, a_T) = \phi_T(s_T, a_T)^{\intercal} \hat{\mathbf{w}}_{T[d]}, \quad \hat{\mathbf{w}}_{T[d]} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_i \left(\phi_T(s_T^i, a_T^i)^{\intercal} \mathbf{w} - r_{T[d]}^i \right)^2 \tag{1}$$

giving the estimated vector-valued expected reward function, which we denote $\hat{\mathbf{Q}}_T(s_T, a_T) = (\hat{Q}_{T[1]}(s_T, a_T), ..., \hat{Q}_{T[D]}(s_T, a_T))^{\intercal}$. Here, $\phi_T(s_T, a_T)$ is a feature vector of state and action. Having obtained the $\hat{\mathbf{Q}}_T$ from (1), we construct an NDP Π_T that gives, for each state, the actions one might take at the last time point. For each state s_T at the last time point, each action a_T is associated with a *unique* vector-valued estimated expected reward given by $\hat{\mathbf{Q}}_T(s_T, a_T)$. Thus, we decide which among these vectors is a desirable outcome using our solution concept, and include their associated actions in $\Pi_T(s_T)$.

For $t < T$, it is only possible to define the expected return of taking an action in a given state by also deciding which particular policy will be followed to choose future actions. In standard fitted-Q, for example, one assumes that the future policy is given by $\pi_j(s) = \arg\max_a \hat{Q}_j(s, a)$ for all $j > t$. In the non-deterministic setting, we may know that the future policy belongs to some set of possible policies derived from Π_j for $j > t$, but in general we do not know which among that set will be chosen; therefore, we explicitly include the dependence of $\hat{\mathbf{Q}}_t$ on the choice of future policies $\pi_j, t < j \leqslant T$ by setting $\hat{\mathbf{Q}}_t(s_t, a_t; \pi_{t+1}, ..., \pi_T) = [\hat{Q}_{t[1]}(s_t, a_t; \pi_{t+1}, ..., \pi_T), ..., \hat{Q}_{t[D]}(s_t, a_t; \pi_{t+1}, ..., \pi_T)]^{\intercal}$ where for $d = 1, \ldots, D$, $\hat{Q}_{t[d]}(s_t, a_t; \pi_{t+1}, ..., \pi_T) = \phi_t(s_t, a_t)^{\intercal} \hat{\mathbf{w}}_{t[d]\pi_{t+1},...,\pi_T}$, and $\hat{\mathbf{w}}_{t[d]\pi_{t+1},...,\pi_T} = \operatorname{argmin}_{\mathbf{w}} \sum_{i=1}^{n} [\phi_t(s_t^i, a_t^i)^{\intercal} \mathbf{w} - \{r_{t[d]}^i + \hat{Q}_{t+1[d]}(s_{t+1}^i, \pi_{t+1}(s_{t+1}^i); \pi_{t+2}, ..., \pi_T)\}]^2$.

We use $\mathcal{Q}_t$ to denote the Q-*pool*, a set of partially-evaluated Q-functions; each member of $\mathcal{Q}_t$ is a function of s_t and a_t only and assumes a particular fixed sequence $\pi_{t+1}, ..., \pi_T$ of future policies. Each one gives the estimated *expected* return for the given state-action pair and future sequence of policies. Our goal in this work is to augment each one with information about the distribution of those returns, because this represents the distribution of outcomes we would expect patients to experience if they follow the given policy.

3.2 Prediction Regions and Tolerance Regions

A *prediction region* $\mathcal{R}^{\alpha}(Y_1, \ldots, Y_n)$ for a data generating process traps the next observation Y_{n+1} with probability $1 - \alpha$:

$$\mathsf{Pr}\left(Y_{n+1} \in \mathcal{R}^{\alpha}(Y_1, \ldots, Y_n)\right) = 1 - \alpha. \tag{2}$$

Conformal prediction (CP) is a methodology for constructing nonparametric prediction regions under mild assumptions Shafer and Vovk (2008). CP methods produce valid prediction regions under any distribution using a given *nonconformity measure.* A nonconformity measure takes a data sample and produces *nonconformity scores* σ_i that measure the "novelty" of each observation with respect to the whole set. For example, we may define nonconformity in terms of regression predictions as $\sigma_i = |y_i - \widehat{y}_i|$ where $\widehat{y}_i = X\widehat{\beta}$ is the ith fitted value. Let F be the data generating distribution for $Y_1, \ldots, Y_n$. Once the nonconformity scores are obtained for all $i = 1 \ldots n$, the nonconformity σ_{n+1} of a hypothetical additional observation Y_{n+1} can be compared to the observed data to obtain a p-value for the null hypothesis $H_0 : Y_{n+1} \sim F$ by $p(Y_{n+1}; Y_1, \ldots, Y_n) = (n+1)^{-1} \sum_{i=1}^{n+1} \mathbf{1}\{\sigma_i \geqslant \sigma_{n+1}\}$. By definition, if $Y_{n+1} \sim F$, $\mathsf{Pr}(p(Y_{n+1}; Y_1, \ldots, Y_n) \leqslant \alpha) = \alpha$. Therefore, the region described by $\mathcal{R}^{\alpha}_{\text{CP}} = \{Y : p(Y; Y_1, \ldots, Y_n) > \alpha\}$ traps Y_{n+1} with probability $1-\alpha$, and is a $1-\alpha$ prediction region. Conformal prediction guarantees the produced prediction is valid under any given nonconformity measure–parametric or nonparametric– even in finite samples. However, the nonconformity measure does influence the size of the region and therefore its usefulness in practice. Lei et al. (2013) propose the use of a kernel density estimate to produce the nonconformity score. We use their approach in our example below and in our subsequent analysis of CATIE.

A $(1 - \alpha)$, γ-content *tolerance region* $\mathcal{R}^{\alpha,\gamma}(Y_1, \ldots, Y_n)$ has the property

$$\mathsf{Pr}(P_F[\mathcal{R}^{\alpha,\gamma}(Y_1, \ldots, Y_n)] \geqslant \gamma) = 1 - \alpha. \tag{3}$$

Note that this is a much stronger probability statement than (2): it says that each interval we construct, with high probability, captures at least γ of the probability mass of the data generating distribution. Several parametric and non-parametric methods are known for one-dimensional tolerance intervals, and Lizotte and Tahmasebi (2017) demonstrate their use for constructing one-dimensional tolerance for the returns of estimated DTRs. Li and Liu (2008) propose a nonparametric method based on order statistics derived from *data depth*, which is a quantity very similar in spirit to a nonconformity measure. Tolerance intervals are then computed by applying one-dimensional techniques to data depth. Although this method produces tolerance regions, which make a stronger probability statement than prediction regions, (3) is shown to hold only asymptotically. Nonetheless, Li and Liu demonstrate that they can achieve good performance in some finite sample settings (i.e. with hundreds of data points and two dimensions).

4. Methods

We present a modular framework for constructing tolerance regions for multiple objectives. The set of optimal policies are found using methods from Lizotte and Laber (2016), and the regions are constructed using methods from Lizotte and Tahmasebi (2017). Algorithm 1 lays out the major steps for constructing regions from policies learned by these methods.

Algorithm 1 Regions for MOMDPs

```
Inputs: Set of $n$ trajectories; region function $\mathcal{R}$; state action pair of interest $(s_t, a_t)$
Output: Region or regions describing the likely observed returns when starting from $s_t$ and taking action $a_t$
  Compute $Q$-pool for timestep $t$ (Lizotte and Laber, 2016)
  by identifying all non-dominated policies
  for each function in the pool and associated policy do
    Collect all trajectories beginning with $a_t$
    For those whose future actions follow the associated policy, retain their observed outcomes
    if there are few such outcomes and we are willing to generalize from trajectories where the policy was not followed then
      Create additional return samples using trajectories that did not follow the policy using residual borrowing (Lizotte and Tahmasebi, 2017)
    else if trajectories were collected using a state-dependent exploration policy then
      Re-weight the empirical distribution of the samples (Lizotte and Tahmasebi, 2017)
    end if
    Adjust the (possibly weighted) returns by regressing them on state and centering the resulting residuals at $Q(s_t, a_t)$
    Construct a region by applying $\mathcal{R}$ to the resulting set of adjusted returns
  end for
  return The regions
```

The methods in Lizotte and Laber (2016) produce a *pool* of Q-functions at each timestep, as described above. For a given a_t, each Q-function in the time t pool produces a different expected return, and its associated future policies produce a different distribution over returns. To construct a region (prediction or tolerance) for a particular Q-function from the pool, we identify the trajectories that start with a_t and whose actions are consistent with that Q-function's assumed future policies. We then use the empirical distribution of their returns to construct whichever type of region is desired. Since in general we do not know which future policy will be followed, we propose to construct regions for all Q functions in the pool and examine their union, their intersection, and the difference of the two. However, other summaries of the regions may also be useful.

Lizotte and Tahmasebi note that if the exploration policy is allowed to depend on state, then the distribution of matched trajectories will not in general match the distribution of returns we would obtain by following the future policies under consideration (Lizotte and Tahmasebi, 2017). Hence, constructing regions naïvely using the matched trajectories will yield incorrect results. They characterize this dependence and present an inverse probability weighting procedure for correcting the distribution before constructing regions. We propose to use the same strategy in the MOMDP setting when necessary. They also propose *residual borrowing*, which uses the residuals between the estimated Q values and the returns among the matched trajectories to infer what the distribution of returns would have been among the unmatched trajectories by combining them with the estimated Q-values. (For further detail please see reference Lizotte and Tahmasebi (2017).) This methodology increases the amount of trajectories we can use, and can also be used within our framework; it relies on two key assumptions. First, it assumes that differences in expected return are fully captured by the current state (the standard MDP assumption.) Second, it assumes that the residual distribution of returns around the expected value is correctly captured by the regression method used. For linear regression, this implies an assumption that the residual distribution is the same across the state space. If there is concern that this is not the case, the correct residual distribution could be estimated by more flexible regression methods, for example heteroscedastic Gaussian process regression (Lzaro-Gredilla and Titsias, 2011).

Lizotte and Tahmasebi assume a discrete state space at the time point of interest. In this work, we accommodate continuous state spaces with function approximation at the first timepoint by extending the residual borrowing idea. Once we we have a return sample for every trajectory that begins with a_t, whether it came from following the policy of interest or was created using residual borrowing, we adjust the values of those returns by regressing them on the current state and re-centering them at the Q-value for the state of interest. Again, this requires an assumption that the regression procedure correctly captures the residual distribution independent of state, as discussed previously. If there is reason to believe this variance is similar across state, however, then this approach allows us to generalize across states at the first timepoint.

Our framework can accommodate any multivariate region method (prediction or tolerance) and any reweighting procedure (based on generalized additive models, density ratio estimation, etc.). In our opinion the most promising methods are the Lei-Robins-Wasserman conformal prediction method (2013) and the Li-Liu tolerance region method (2008).

5. Results

Below, we review our goals and present results of applying our method to the CATIE data.

5.1 Evaluation Goals

Our goal is to illustrate the effect that sequential decision-making, in the form of different choices of future policy, can have on the likely outcomes associated with different immediate actions. Thus, rather than comparing a numerical measure of success, we will demonstrate visually the additional structure that we are able to extract from the data to aid decision-making. Note that previous work (Lizotte and Tahmasebi, 2017) has investigated the statistical properties of these methods, in terms of width and coverage; this was done using synthetic data. For this work, since we do not have a ground truth generative model, we have chosen the methods that are most conservative from those investigated by Lizotte and Tahmasebi (2017).

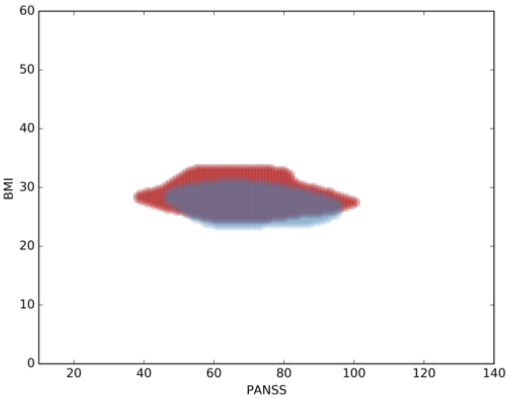

Figure 2: Conformal prediction region for outcomes with $\alpha = 0.2$, assuming exploration policy at second stage

Figure 2 illustrates what is possible using off-the-shelf conformal prediction method of Lei et al. (2013). It show two prediction regions with level $\alpha = 0.2$, one for ziprasidone and one for olanzapine, for a participant with a baseline PANSS of 75 and a BMI of 27.2, which is near the median of the population. These regions are constructed using the observed outcomes from the study. Hence, these are the regions that would be expected to trap one additional outcome *if a patient followed the policy used to collect the data*, that is, if the patient followed a random policy at the second stage. We can see that the region for olanzapine is a bit higher in the BMI direction, reflecting its propensity to induce weight gain. It is also shifted somewhat to the left also extends a bit further to the left and right on the PANSS axis, reflecting that although olanzapine appears to produce lower PANSS *on average*, in the study outcomes were more variable than they were among patients who started with ziprasidone.

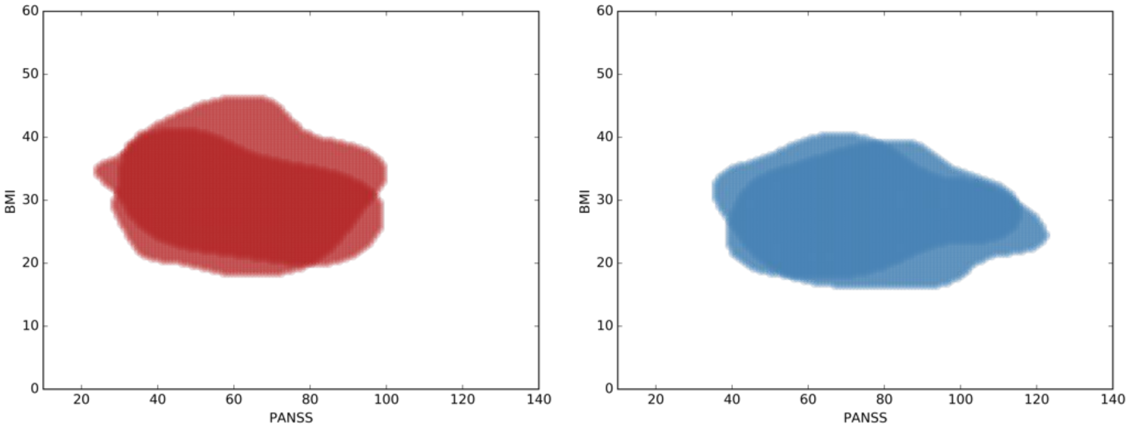

Figure 3: Two example regions for olanzapine and ziprasidone, conditioned on an initial state of PANSS = 75 and BMI = 27.2, and conditioned on specific (deterministic) future policies.

5.2 Prediction Regions With Sequential Decision-Making

Using the methods of Lizotte and Laber (2016), we computed over 29,000 possible future policies for the second stage of CATIE for each of olanzapine and ziprasidone. Each of these policies were non-dominated, using an indifference region of 15 points of PANSS and 1.5 points of BMI. For each of these policies, we computed a conformal prediction region. Figure 3 shows two randomly selected regions for each of olanzapine and ziprasidone at the first stage. The different shapes of the regions for the same action reflects the ability to influence how outcomes are distributed through the phase 2 policy.

To summarize these tens of thousands of regions, we present three plots shown in Figure 4. The leftmost plot shows the intersection of all of the regions. These show outcomes that are contained in the 80% prediction region regardless of what future policy is chosen. These represent likely outcomes that are in a sense unavoidable. The centre plot shows the union of all the regions. These show that are in the 80% prediction region for at least one

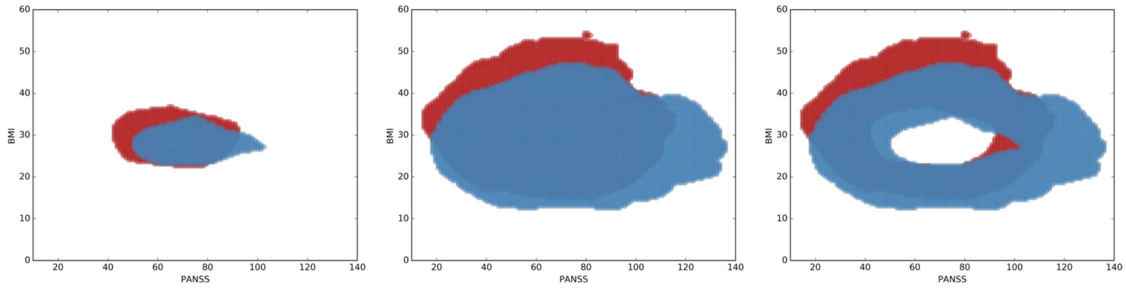

Figure 4: Summaries of all prediction regions for olanzapine and ziprasidone, conditioned on initial state of of PANSS = 75 and BMI = 27.2. Left: intersection of all regions. Centre: union of all regions. Right: union minus intersection.

future policy. The rightmost plot shows the difference between the union and the intersection. Each point here is *excluded* from the 80% prediction region by at least one policy, and is *included* in the 80% prediction region by at least once policy. Hence, this "ring" of points represents the area of outcome space that we have some control over after choosing the first action.

6. Discussion

Our results show how one might present and use the information from the regions: by overlaying the regions corresponding to different actions, we can compare their likely outcomes as an aid for decision-making. We expect that the best choices for the details of this framework (e.g. approximation method, region type, presentation style) will differ from task to task, and it is therefore our hope that this work provides a foundation for both novel research and useful application. We envision our method forming the underpinnings of an interactive piece of software that allows the user to explore the interaction between choice of future policy and likely outcomes. For example, we could allow a user to click on a point in the difference plot (rightmost plot of Figure 4) to identify the policies that either include or exclude that point. More generally, we could have the user "paint" regions of the outcome space to include or exclude, in order to help identify the most appropriate future policies according to their preferences.

There are two main methodological limitations of the current work that we feel are most important to address going forward. First, we note that the regions, for example in Figure 3, tend to be larger than those constructed for the exploration policy. This may seem counterintuitive, but consider that when constructing a region for a learned policy, we actually have *less* data that exactly follow that policy than there are trajectories in the entire dataset. We attempt to mitigate this with residual borrowing, but it may be that our estimates of the residual distribution are wider than they need to be. This is a focus in our ongoing work. Second, although in this illustrative example we used a complete-case analysis, going forward it will be important to incorporate uncertainty information, for example from multiple imputations, to avoid bias from study dropout. This may lead to Bayesian formulations of the region problem as a whole that could provide for better incorporation of prior information.

References

D. B. Allison, J. L. Mentore, M. Heo, L. P. Chandler, J. C. Cappelleri, M. C. Infante, and P. J. Weiden. Antipsychotic-induced weight gain: A comprehensive research synthesis. *American Journal of Psychiatry*, 156:1686–1696, November 1999.

D. P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. II.* Athena Scientific, 3rd edition, 2007. ISBN 1886529302, 9781886529304.

S. R. Kay, A. Fiszbein, and L. A. Opfer. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, 13(2):261–276, 1987.

E. B. Laber, D. J. Lizotte, M. Qian, W. E. Pelham, and S. A. Murphy. Dynamic treatment regimes: technical challenges and applications. *Electronic Journal of Statistics*, 8(1): 1225–1272, 2014.

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Stefan Leucht, Andrea Cipriani, Loukia Spineli, Dimitris Mavridis, Deniz Örey, Franziska Richter, Myrto Samara, Corrado Barbui, Rolf R. Engel, John R. Geddes, Werner Kissling, Marko Paul Stapf, Bettina Lässig, Georgia Salanti, and John M. Davis. Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis. *The Lancet*, 382(9896):951–962, June 2013. ISSN 0140-6736. doi: 10.1016/S0140-6736(13)60733-3. URL http://dx.doi.org/10.1016/S0140-6736(13)60733-3.

Jun Li and Regina Y. Liu. Multivariate spacings based on data depth: I. construction of nonparametric multivariate tolerance regions. *The Annals of Statistics*, 36(3):1299–1323, 06 2008. doi: 10.1214/07-AOS505. URL http://dx.doi.org/10.1214/07-AOS505.

D. J. Lizotte, M. Bowling, and S. A. Murphy. Linear fitted-Q iteration with multiple reward functions. *Journal of Machine Learning Research*, 13:3253–3295, Nov 2012.

Daniel J. Lizotte and Eric B. Laber. Multi-objective markov decision processes for data-driven decision support. *Journal of Machine Learning Research*, 17(211):1–28, 2016. URL http://jmlr.org/papers/v17/15-252.html.

Daniel J. Lizotte and Arezoo Tahmasebi. On prediction and tolerance intervals for Dynamic Treatment Regimes. *arXiv*, 2017. 1704.00000.

M. Lzaro-Gredilla and M.K. Titsias. Variational heteroscedastic gaussian process regression. In *28th International Conference on Machine Learning (ICML)*, 2011.

M. Milani Fard and J. Pineau. Non-deterministic policies in Markovian decision processes. *Journal of Artificial Intelligence Research*, 40:1–24, 2011.

L. Orellana, A. Rotnitzky, and J. Robins. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: Main content. *Int. Jrn. of Biostatistics*, 6(2), 2010.

Miquel Porta, editor. *A Dictionary of Epidemiology.* Oxford University Press, 2014.

D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, March 2008.

S. Shortreed, E. B. Laber, D. J. Lizotte, T. S. Stroup, J. Pineau, and S. A. Murphy. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine Learning*, 84(1–2):109–136, 2011.

T. S. Stroup and al. The national institute of mental health clinical antipsychotic trials of intervention effectiveness (CATIE) project: Schizophrenia trial design and protocol development. *Schizophrenia Bulletin*, 29(1), 2003.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, 1998.

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

Appendix A.

CATIE was an 18-month study of $n = 1460$ patients that was divided into two main phases of treatment. Upon entry, most patients began "Phase 1," and were randomized to one of five treatments with equal probability: olanzapine◂, risperidone▸, quetiapine▴, ziprasidone▾, or perphenazine■. As time passed, patients were given the opportunity to discontinue their Phase 1 treatment and begin "Phase 2" on a new treatment. The possible Phase 2 treatments depended on the reason for discontinuing Phase 1 treatment. If the Phase 1 treatment was ineffective at reducing symptoms, then patients entered the "Efficacy" arm of Phase 2, and their Phase 2 treatment was chosen randomly as: {clozapine◆} with probability 1/2, or uniformly randomly from the set {olanzapine◂, risperidone▸, quetiapine▴} with probability 1/2. Because relatively few patients entered this arm, and because of the uneven action probabilities, it is reasonable to combine {olanzapine◂, risperidone▸, quetiapine▴} into one "not-clozapine" action, and we will do so here. If the Phase 1 treatment produced unacceptable side-effects, they entered the "Tolerability" arm of Phase 2, and their Phase 2 treatment was chosen uniformly randomly from {olanzapine◂, risperidone▸, quetiapine▴, ziprasidone▾}.

Basis Rewards

We use ordinary least squares to learn Q functions for two basis rewards. For our first basis reward, we use the Positive and Negative Syndrome Scale (PANSS) which is a numerical representation of the severity of psychotic symptoms experienced by a patient (Kay et al., 1987). PANSS has been used in previous work on the CATIE study (Shortreed et al., 2011; Lizotte et al., 2012; Swartz et al., 2003), and is measured for each patient at the beginning of the study and at several times over the course of the study. Larger PANSS scores are worse, so we minimize rather than maximize when learning policies.

For our second basis reward, we use Body Mass Index (BMI), a measure of obesity. Weight gain is an important and problematic side-effect of many antipsychotic drugs (Allison et al., 1999), and has been studied in the multiple-reward context (Lizotte et al., 2012). Because having a larger BMI is worse in this population, again we minimize rather than maximize when learning policies.

State Space

For our state space, we use the patient's most recently recorded PANSS score, which experts consider for decision making (Shortreed et al., 2011). We also include their most recent BMI, and several baseline characteristics.

Because the patients who entered Phase 2 had different possible action sets based on whether they entered the Tolerability or Efficacy arm, we learn separate Q-functions for these two cases. The feature vectors we use for Stage 2 Efficacy patients are given by

$$\begin{aligned}\phi^{\text{EFF}}(s_2, a_2) = \ &[1,\ 1_{\text{TD}},\ 1_{\text{EX}},\ 1_{\text{ST1}},\ 1_{\text{ST2}},\ 1_{\text{ST3}},\ 1_{\text{ST4}},\ s_{2:\text{P}},\ s_{2:\text{B}}, \\ &\qquad\qquad\qquad\qquad 1_{a_2=\blacklozenge},\ s_{2:\text{P}} \cdot 1_{a_2=\blacklozenge},\ s_{2:\text{B}} \cdot 1_{a_2=\blacklozenge}]^\intercal.\end{aligned}$$

Here, $s_{2:\text{P}}$ and $s_{2:\text{B}}$ are the PANSS and BMI percentiles at entry to Phase 2, respectively. Feature $1_{a_2=\blacklozenge}$ indicates that the action at the second stage was clozapine◆ and not one of the other treatments. We also have other features that do not influence the optimal action choice but that are chosen by experts to reduce variance in the value estimates.[2] 1_{TD} indicates whether the patient has had tardive dyskinesia (a motor-control side-effect), 1_{EX} indicates whether the patient has been recently hospitalized, and 1_{ST1} through 1_{ST4} indicate the "site type," which is the type of facility at which the patient is being treated (e.g. hospital, specialist clinic, etc.)

For Phase 2 patients in the Tolerability arm, the possible actions are $\mathcal{A}_2^{\mathrm{TOL}} = \{\blacktriangleleft, \blacktriangle, \blacktriangleright, \blacktriangledown\}$, and the feature vectors we use are given by

$$\begin{aligned}\phi^{\mathrm{TOL}}(s_2, a_2) = \ [&1,\ 1_{\mathrm{TD}},\ 1_{\mathrm{EX}},\ 1_{\mathrm{ST1}},\ 1_{\mathrm{ST2}},\ 1_{\mathrm{ST3}},\ 1_{\mathrm{ST4}},\ s_{2:\mathrm{P}},\ s_{2:\mathrm{B}},\\ &1_{a_2=\blacktriangleleft},\ s_{2:\mathrm{P}}\cdot 1_{a_2=\blacktriangleleft},\ s_{2:\mathrm{B}}\cdot 1_{a_2=\blacktriangleleft},\ 1_{a_2=\blacktriangle},\ s_{2:\mathrm{P}}\cdot 1_{a_2=\blacktriangle},\ s_{2:\mathrm{B}}\cdot 1_{a_2=\blacktriangle},\\ &1_{a_2=\blacktriangleright},\ s_{2:\mathrm{P}}\cdot 1_{a_2=\blacktriangleright},\ s_{2:\mathrm{B}}\cdot 1_{a_2=\blacktriangleright}]^{\intercal}.\end{aligned}$$

Here we have three indicator features for different treatments at Phase 2, $1_{a_2=\blacktriangleleft}$, $1_{a_2=\blacktriangleright}$, $1_{a_2=\blacktriangle}$, with ziprasidone represented by turning all of these indicators off. Again we include the product of each of these indicators with the PANSS percentile s_2. The remainder of the features are the same as for the Phase 2 Efficacy patients.

For Phase 1 patients, the possible actions are $\mathcal{A}_1 = \{\blacktriangleleft, \blacksquare, \blacktriangle, \blacktriangleright, \blacktriangledown\}$, and the feature vectors we use are given by

$$\begin{aligned}\phi^{\mathrm{EFF}}(s_2, a_2) = \ [&1,\ 1_{\mathrm{TD}},\ 1_{\mathrm{EX}},\ 1_{\mathrm{ST1}},\ 1_{\mathrm{ST2}},\ 1_{\mathrm{ST3}},\ 1_{\mathrm{ST4}},\ s_{1:\mathrm{P}},\ s_{1:\mathrm{B}},\\ &1_{a_2=\blacktriangleleft},\ s_{1:\mathrm{P}}\cdot 1_{a_2=\blacktriangleleft},\ s_{1:\mathrm{B}}\cdot 1_{a_2=\blacktriangleleft},\ 1_{a_2=\blacksquare},\ s_{1:\mathrm{P}}\cdot 1_{a_2=\blacksquare},\ s_{1:\mathrm{B}}\cdot 1_{a_2=\blacksquare},\\ &1_{a_2=\blacktriangle},\ s_{1:\mathrm{P}}\cdot 1_{a_2=\blacktriangle},\ s_{1:\mathrm{B}}\cdot 1_{a_2=\blacktriangle},\ 1_{a_2=\blacktriangleright},\ s_{1:\mathrm{P}}\cdot 1_{a_2=\blacktriangleright},\ s_{1:\mathrm{B}}\cdot 1_{a_2=\blacktriangleright}]^{\intercal}.\end{aligned}$$

We have four indicator features for different treatments at Phase 2, $1_{a_1=\blacktriangleleft}$, $1_{a_1=\blacksquare}$, $1_{a_1=\blacktriangle}$, and $1_{a_1=\blacktriangleright}$, with ziprasidone represented by turning all of these indicators off. We include the product of each of these indicators with the PANSS percentile s_1 at entry to the study, and the remainder of the features are the same as for the Phase 2 feature vectors. (These are collected before the study begins and are therefore available at Phase 1 as well.)

[2] See Section 4.2 of the paper by Shortreed et al. (2011) for an explanation of these kinds of features.