Using Causal Inference to Estimate What-if Outcomes for Targeting Treatments

Qing Liu[*1] **Katharine Henry**[*1] **Yanbo Xu**[1] **Suchi Saria**[1,2]

[1]**Department of Computer Science**
[2]Department of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD 21218
{ qingliu, keh, yxu68, ssaria} @jhu.edu

1 Introduction

Individuals often have heterogeneous outcomes after interventions. As a result, clinicians constantly ask themselves, given a patient's history, what would happen to the patient's clinical trajectory if they were given one treatment versus another. In order to target care, estimating how outcomes or responses to treatments will vary across individuals is critical. However, in practice it is often unknown how the patient's signals will change in response to a treatment until that treatment is actually administered. Furthermore, even if we have observed data on the patient for one treatment, the *counterfactual*, i.e., what would have happened to the patient if the doctor had made a different choice, is unobserved.

In this paper, we frame this as a causal question and use two different Bayesian nonparametric (BNP) approaches to estimate the patient's clinical trajectory in response to treatments. As a motivating application, we focus on modeling blood pressure (BP) and heart rate (HR) for patients in the intensive care unit (ICU) and estimate their responses to six types of treatments that are used in their management. These two signals are among the most commonly used vital signs in the ICU and are critical for identifying life-threatening conditions like septic and hemorrhagic shock [Dellinger et al., 2013]. We begin by describing the causal question and the assumptions under which the outcomes are consistently estimated. We then provide a comparison of two state-of-the-art estimation procedures on the BP and HR trajectory estimation task.

2 Approaches

To estimate what would happen to an individual's trajectory under different interventions, we first frame the causal what-if problem as a statistical one. We build upon the framework proposed for estimating individual-specific treatment responses and the resulting clinical trajectories as discussed in Xu et al. [2016].

Assume we have observations $\boldsymbol{Y}_i = \{Y_{i1}, ..., Y_{iJ_i}\}$ that were irregularly-measured at times $\boldsymbol{t}_i = \{t_{i1}, ..., t_{iJ_i}\}$, treatments $\boldsymbol{A_i} = \{A_{i1}, ..., A_{iL_i}\}$ that were irregularly-prescribed at times $\boldsymbol{\tau_i} = \{\tau_{i1}, ..., \tau_{iL_i}\}$, and a set of covariates $\boldsymbol{C}_{ij} \in \mathbb{R}^p$ at time t_{ij} that could possibly affect both the measurement and treatment assignment, then the causal question we want to ask is: at time t_{ij}, given the treatments history $\boldsymbol{a}_{i,\leq j-1}$, measurements history $\boldsymbol{y}_{i,\leq j-1}$, and covariates $\boldsymbol{c}_{ij}$, what would the measurement have been if we had added a new treatment a_{ij} for the ith individual? To answer this question, we adopt Neyman-Rubin's framework of *potential outcome* [Rubin, 2011] to account for *conterfactuals* that are not observed from the data. This gives us a formal mathematical representation of the potential outcome using the notation $\mathbb{E}[Y(a)]$, which reads as the expectation of Y if we *do* a. We treat $\mathbb{E}[Y(a)]$ as a random variable and posit a distribution to model its value as a function of clinical events.

[*]The authors contributed equally to this work.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Next, to estimate the potential outcomes, we use the *g-formula* [Robins, 1986]. Under the following two assumptions—*Consistency* (the potential outcome is the same as the observed outcome under the treatments that are actually assigned to the patient) and *Sequential Ignorability* (the treatment assignment is independent of the future potential outcomes given the past treatments, measurements, and covariates)—we can build a link between the potential outcomes and the observational data using g-formula. As a result, we can formulate the problem of estimating the ith individual's potential outcome at time t_{ij} as follows:

$$\mathbb{E}[Y_{ij}(a_{ij}) \mid \boldsymbol{a}_{i,\leq j-1}, \boldsymbol{y}_{i,\leq j-1}, \boldsymbol{c}_{ij}] = \mathbb{E}[Y_{ij} \mid \boldsymbol{A}_{ij} = a_{ij}, \boldsymbol{a}_{i,\leq j-1}, \boldsymbol{y}_{i,\leq j-1}, \boldsymbol{c}_{ij}]. \tag{1}$$

Since the true underlying model for the conditional expectation of the measurements in Eq.1 is unknown, we need a flexible means for estimating this from data. The formulation of the estimation task is driven by three properties of our data. First, the measurements are often sparse and irregularly-sampled. Second, we want to estimate a continuous clinical trajectory rather than a single point-in-time estimate. Finally, there is large variation in response to treatment across individuals, which necessitates estimating patient-specific response to treatment interventions rather than only estimating average population-level effects.

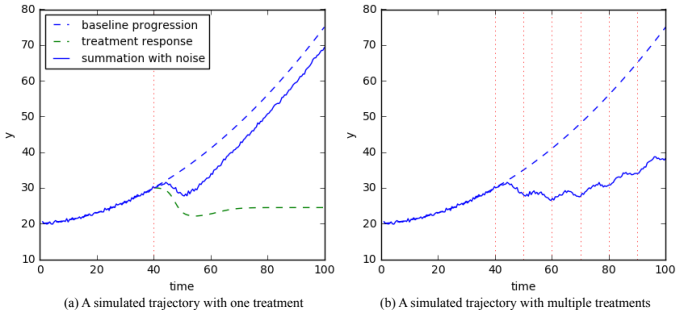
To address these challenges, we first consider a BNP method for estimating individualized treatment response (ITR) that was recently proposed by Xu et al. [2016]. ITR posits a functional form to model the continuous trajectory of signals under treatments. This allows it to both make continuous predictions into the future and to learn from sparse, irregularly-sampled signals without ad-hoc processing of the data. ITR also accounts for individual variation in treatment response by learning a patient-specific posterior distribution of the parameters based on the individual's past responses. In comparison, standard approaches for modeling heterogeneous treatment responses rely on defining subpopulations from the data and estimating the average effect for each subpopulation given the covariates [Foster et al., 2010, Imai et al., 2013, Tian et al., 2014, Athey and Imbens, 2015]. Of these approaches, Bayesian Additive Regression Trees (BART) [Chipman et al., 2010] has emerged as a popular method for estimating treatment response. It has been successfully used in a number of causal inference tasks [Green and Kern, 2010, Hill, 2011] and was one of the winning approaches at the 2016 Atlantic Causal Inference Conference Competition [Dorie et al., 2016] for estimating treatment response. Due to the recent success of BART, we compare its performance in our setting to that of ITR. Since BART only gives a point in time estimate, we adapt it to our problem setting by evaluating the model at a sequence of discrete time intervals, which we explain in details in the following section.

ITR: The nonparametric Bayesian approach for estimating individualized treatment-response curves (ITR) [Xu et al., 2016] assumes that the measurements are a sum of three components and can be written as

$$\boldsymbol{y}_i \mid \boldsymbol{a}_i, \boldsymbol{c}_i = \underbrace{u_i(\boldsymbol{c}_i)}_{\text{baseline progression}} + \underbrace{f_i(\boldsymbol{a}_i)}_{\text{treatment responses}} + \underbrace{\boldsymbol{\epsilon}_i}_{\text{noise}}, \tag{2}$$

where u_i models the patient's baseline progression if no treatment was prescribed, f_i models the cumulative response to sequential prescriptions of treatments, and ϵ_i models the i.i.d. noises at each time point. Here, we assume that u_i follows a Gaussian Process with a mean function defined by a linear regression over the covariates and a time-dependent covariance term in the form of a squared exponential. Alternate forms for u_i that incorporate long and short term structure in the baseline disease trajectories [Schulam and Saria, 2015] should be incorporated when relevant. ITR also assumes treatment effects are additive by modeling f_i as a summation of treatment-specific response curves as illustrated in Figure 1. Finally, ITR posits a Dirichlet process mixture (DPM) prior on each individual's parameters to automatically learn the sub-structure in the population and allow shared statistical strength in the learning process across individuals.

BART: Bayesian Additive Regression Trees (BART) has been successfully used in a variety of tasks including drug discovery [Chipman et al., 2010] and estimating treatment effect on cognitive scores of low-birth-weight infants [Hill, 2011]. It flexibly models the measurement $\boldsymbol{Y}$ with a nonparametric Bayesian approach that uses a sum-of-trees with a regularization prior. While BART provides an estimate of the measurement at a single point, in our setting, we need a continuous response over time. To achieve this, we divide the time window into multiple intervals and assume that the trajectory is a piecewise-constant function where the value for each interval is predicted using a separate BART instance. For example, to predict the signal values for the future 36 hours, we train six BART models,



(a) A simulated trajectory with one treatment (b) A simulated trajectory with multiple treatments

Figure 1: An illustration of additive treatment responses over time

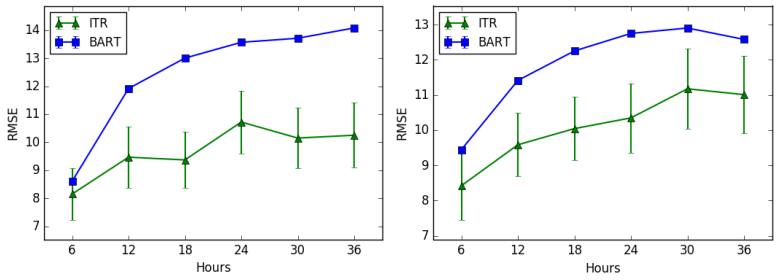
one for each six hour interval from 0 to 36 hours into the future. Each instance uses the observed measurements thus far to predict the value of the trajectory within the specified time interval.

3 Experimental Results

We apply the approaches described above to our task of estimating HR and BP clinical trajectories from electronic health record (EHR) data. We assess the methods based on the accuracy of the predicted clinical trajectories and a qualitative comparison of the what-if outcome under different treatments. We also provide a qualitative example of the utility of the individualized treatment response curves learned by ITR.

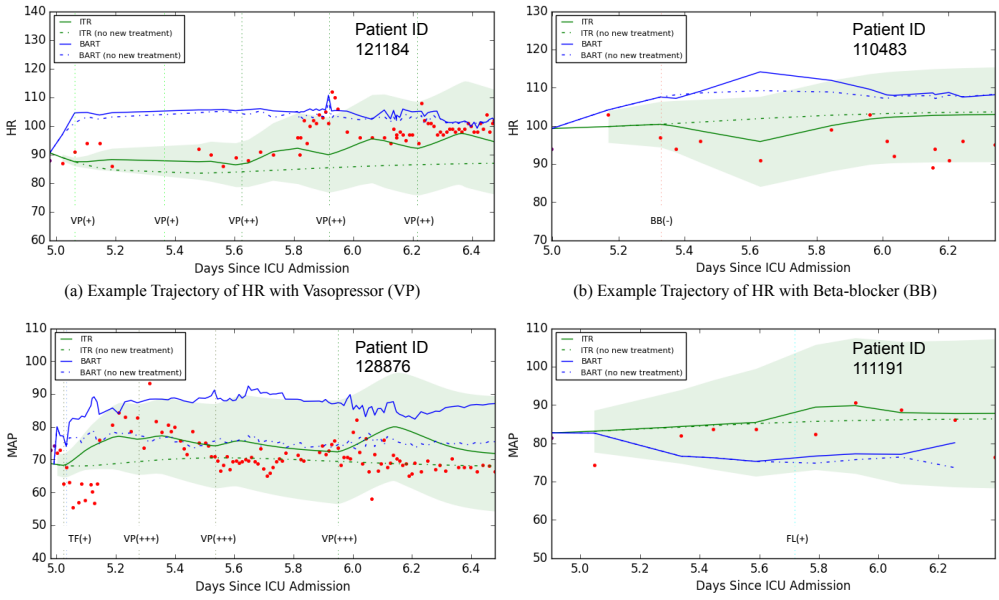
Datasets: EHRs were collected over two years from patients who entered the ICU at Howard County General Hospital. In this study, we focus on modeling the longitudinal measurements of heart rate (HR) and mean arterial pressure (MAP), a type of blood pressure, and estimating individual responses to 6 typical treatments for managing MAP and HR: vasopressors, beta-blockers, ACE inhibitors, vasodilators, blood transfusions, and fluid boluses. The EHR data contains 900 ICU patients who were prescribed at least one of the treatments. For each patient we extract chronic conditions (age, weight, chronic airway obstruction, chronic bronchitis, chronic heart failure, chronic kidney, chronic liver disease, chronic pancreatitis, chronic pulmonary, diabetes, emphysema, end-stage renal disease, gender, hematologic malignancy, immunodeficiency, and metastatic carcinoma) and acute events (acute organ failure, qSOFA, severe acidosis, systemic inflammatory response syndrome (SIRS), acute pancreatitis, acute liver failure, acute respiratory distress syndrome (ARDS), myocardial infarction, obstructive shock, peritonitis, pneumonia, sepsis, urinary tract infection, and hemorrhage). We use at most one day prior to ICU admission until up to 4 days after ICU admission as training data and use an additional 36 hours as test data. We randomly select a subset of 300 patients to include in the study. The training set contains $27,515$ measurements of MAP, $39,522$ measurement of HR, and $2,237$ treatment instances.

Results: To quantitatively evaluate the model, we individually predict the patient's MAP and HR given the treatments prescribed in the EHR data for 36 hours after the training period and compare the predicted values with the observed measurements. In Figure 2, we report the average root-mean-square error (RMSE) over all patients as a function of the number of hours since the last observation used to train the model. The error bars show the 95% confidence intervals. In both signals and all prediction time frames, ITR consistently achieves smaller prediction error compared with BART.



(a) Prediction errors on heart rate (HR) (b) Prediction errors on mean arterial pressure (MAP)

Figure 2: Root mean squared errors (RMSE) for predicting MAP and HR in the future 36 hours



(a) Example Trajectory of HR with Vasopressor (VP)

(b) Example Trajectory of HR with Beta-blocker (BB)

(c) Example Trajectory of MAP with Vasopressor (VP) and Transfusion (TF)

(d) Example Trajectory of MAP with Fluid Bolus (FL)

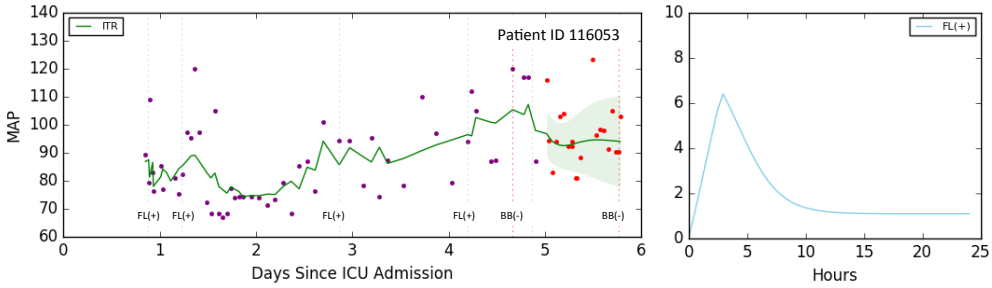
Figure 3: Example prediction trajectories of HR and MAP

We further make qualitative evaluations by looking into example trajectories of patients and comparing the predicted response under different treatment regimes. In Figure 3 we plot the predictions made by ITR (in green) and BART (in blue). Red dots represent test data and the vertical dashed lines show timing of different treatments with different colors. The solid lines represent the predictions made by the two models with the observed treatments, while the dashed lines predict what the value would be if no new treatments were added (the counterfactual). In patients 121184 and 128876 we see that both methods show the expected clinical outcome, namely that vasopressor use increases both MAP and HR compared to the counterfactual. However, ITR shows a more pronounced treatment response and more closely matches the observed data under treatment. In patient 110483, ITR captures the expected effect of beta-blocker use decreasing heart rate, while BART shows the opposite trend.

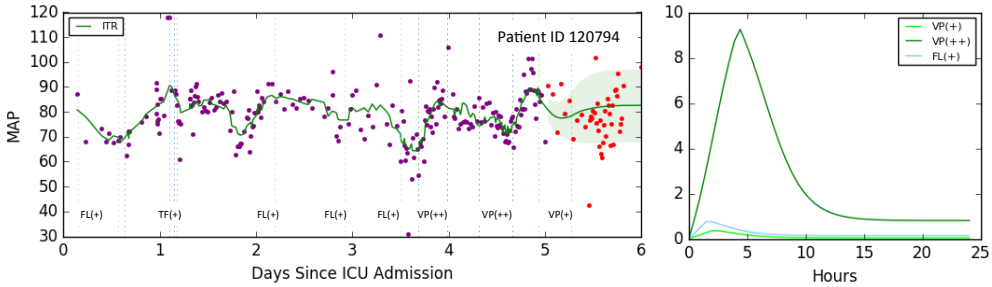
To demonstrate the potential utility of the individualized treatment responses learned by ITR, consider the two patients presented in Figure 4. As in Figure 3, the dots represent observed data and vertical dashed lines represent treatments. Purple dots represent samples used to estimate the treatment effect and red dots represent samples used in testing. The outcome estimated by ITR is shown in green. To the right, we show the estimated treatment responses learned by ITR for fluids and vasopressors (if prescribed). The patients depicted share many characteristics: both are male, of similar age, and immunosuppressed, and both experienced hypotension (MAP $<$ 65 mmHg) related to sepsis. For patient A, the ITR estimate shows a large increase in blood pressure after administering a fluid bolus, and in fact when the patient became hypotensive, he received several fluid boluses and his blood pressure increased and stabilized. In contrast, the estimated responses for patient B show little response to fluids or low doses of vasopressors, but a good response to a mid-level dose of vasopressors. As shown in Figure 4b, Patient B was managed for days with aggressive fluid resuscitation, but his blood pressure kept dropping despite the fluid boluses. At the end of the third day in the ICU he became so hemodynamically unstable that he required a higher level of therapy. Vasopressors were subsequently started at a mid-level dose and his blood pressure stabilized. In this case, seeing from the ITR curves that patient B has limited response to fluids may have prompted the clinician to seek more aggressive therapies earlier and potentially have improved care.

4 Conclusion

Clinically, it is often challenging to know how a patient will respond to a given treatment and what would happen if that patient was given a different treatment. While other works have looked at estimating treatment response, we specifically consider the case of predicting response over time from sparse, irregularly-sampled EHR data. We use g-formula to map the causal what-if question to



(a) Example Trajectory of MAP and Treatment Response Curve of Fluid (FL)



(b) Example Trajectory of MAP and Treatment Response Curves of Vasopressor (VP) and Fluid (FL)

Figure 4: Example trajectories of MAP and treatment response curves

a statistical problem and experiment with two BNP approaches. We compare their performance at estimating blood pressure and heart rate responses to a group of six treatments.

Both ITR and BART flexibly model individual signal responses and achieve good prediction error in our problem setting. Compared to ITR, BART is easier to implement and to apply to unseen patients. Nevertheless, ITR has several key advantages over BART. First, ITR is designed to learn individualized parameters to describe the treatment response curve, while BART models variation in treatment response by conditioning on the observed covariates. BART is only able to capture differences between populations that can be defined from the observed covariates. However, in medicine, it is frequently unknown what precisely differentiates a responder from a non-responder for a given treatment, as in the case of Figure 4. Second, since ITR models the entire history and learns a trajectory rather than a point in time estimate, this allows it to make continuous predictions without having to make ad-hoc decisions about how to represent the sparse, irregularly-sampled data. In contrast our adapted version of BART only makes predictions at discrete time intervals and requires summarizing the past in a fixed time window. Finally, the hierarchical prior in the ITR model allows it to share statistical strength across the population and also to incorporate domain knowledge. While BART uses a regularization prior to prevent overfitting, this prior is not typically used to introduce domain knowledge. Overall, the results suggests that ITR consistently achieves better prediction accuracy, especially at points further into the future. Moreover, qualitative assessment shows that the estimates from ITR more closely match clinical expectations.

As in any causal model with observational data, these methods rely on several assumptions. To apply the g-formula, we assume consistency and sequential ignorability. Any unobserved confounding will introduce bias to the model. This requires thoughtful consideration of the problem in order to limit confounding. Moreover, since the underlying model of the conditional expectation of the measurements is unknown, any estimator could suffer from model misspecification. Both methods address this problem by using BNP to avoid specifying a particular parametric form.

Estimating individualized what-if outcomes has exciting clinical implications. The ability to answer what would happen to a patient if different treatments were prescribed before actually administering the treatment would allow clinicians to customize interventions based on expected response.x We show that BNP methods can be used to flexibly estimate what-if outcomes and potentially provide clinical support.

References

Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050:5, 2015.

Hugh A Chipman, Edward I George, and Robert E McCulloch. BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1):266–298, 2010.

RP Dellinger, MM Levy, A Rhodes, and et al. Surviving sepsis campaign: International guidelines for management of severe sepsis and septic shock. *Critical Care Medicine*, 41(2):580–637, 2013.

Vincent Dorie, Jennifer Hill, Uri Shalit, Dan Cervone, and Marc Scott. Is your satt where it's at? a causal inference data analysis challenge. In *2016 Atlantic Causal Inference Conference*, 2016.

J Foster, J Taylor, and S Ruberg. Subgroup identification from randomized clinical data. *Statistics in Medicine*, 30:2867–2880, 2010.

Donald P Green and Holger L Kern. Modeling heterogeneous treatment effects in large-scale experiments using bayesian additive regression trees. In *The annual summer meeting of the society of political methodology*, 2010.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Studies*, 20(1):217–240, 2011.

Kosuke Imai, Marc Ratkovic, et al. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

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

Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 2011.

P.F. Schulam and S. Saria. A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In *Advances in Neural Information Processing Systems*, pages 748–756, 2015.

Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109 (508):1517–1532, 2014.

Yanbo Xu, Yanxun Xu, and Suchi Saria. A non-parametric bayesian approach for estimating treatment-response curves from sparse time series. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56. JMLR W&CP, 2016.