Bayesian Sensitivity Analysis for Offline Policy Evaluation

Jongbin Jung
jongbin@stanford.edu
Stanford University

Ravi Shroff
ravi.shroff@nyu.edu
New York University

Avi Feller
UC-Berkeley
afeller@berkeley.edu

Sharad Goel
scgoel@stanford.edu
Stanford University

ABSTRACT

On a variety of complex decision-making tasks, from doctors prescribing treatment to judges setting bail, machine learning algorithms have been shown to outperform expert human judgments. One complication, however, is that it is often difficult to anticipate the effects of algorithmic policies prior to deployment, as one generally cannot use historical data to directly observe what would have happened had the actions recommended by the algorithm been taken. A common strategy is to model potential outcomes for alternative decisions assuming that there are no unmeasured confounders (i.e., to assume *ignorability*). But if this ignorability assumption is violated, the predicted and actual effects of an algorithmic policy can diverge sharply. In this paper we present a flexible Bayesian approach to gauge the sensitivity of predicted policy outcomes to unmeasured confounders. In particular, and in contrast to past work, our modeling framework easily enables confounders to vary with the observed covariates. We demonstrate the efficacy of our method on a large dataset of judicial actions, in which one must decide whether defendants awaiting trial should be required to pay bail or can be released without payment.

CCS CONCEPTS

• **Applied computing** → **Sociology**; • **Computing methodologies** → *Machine learning*.

KEYWORDS

offline policy evaluation, pretrial risk assessment, sensitivity to unmeasured confounding

ACM Reference Format:
Jongbin Jung, Ravi Shroff, Avi Feller, and Sharad Goel. 2020. Bayesian Sensitivity Analysis for Offline Policy Evaluation. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375822

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02…$15.00
https://doi.org/10.1145/3375627.3375822

1 INTRODUCTION

Machine learning algorithms are increasingly used by employers, judges, lenders, and other experts to guide high-stakes decisions [2, 3, 5, 6, 22]. These algorithms, for example, can be used to help determine which job candidates are interviewed, which defendants are required to pay bail, and which loan applicants are granted credit. When determining whether to implement a proposed policy, it is important to anticipate its likely effects using only historical data, as ex-post evaluation may be expensive or otherwise impractical. This general problem is known as offline policy evaluation. One approach to this problem is to first assume that treatment assignment is ignorable given the observed covariates, after which a variety of modeling techniques are theoretically justified, including regression, matching, and doubly robust estimation [1, 10, 23, 24].

As the assumption of ignorability is often unrealistic in practice, sensitivity analysis methods seek to gauge the effects of unmeasured confounding on predicted outcomes. The literature on sensitivity analysis dates back at least to the work of Cornfield et al. [8] on the link between smoking and lung cancer. In a seminal paper, Rosenbaum and Rubin [21] proposed a framework for sensitivity analysis in which a binary unmeasured confounder affects both a binary response variable and a binary treatment decision. Some extensions of the Rosenbaum and Rubin approach to sensitivity analysis include allowing for non-binary response variables, and incorporating machine learning methods into the estimation process [4, 9, 15].

A complementary line of work extends classical sensitivity analysis by taking a Bayesian perspective and averaging, rather than sweeping, over values of the sensitivity parameters [19, 20]. In this setup, even a weakly informative prior distribution over the sensitivity parameters can exclude estimates that are not supported by the data. Several authors have recently extended traditional methods for sensitivity analysis to offline policy evaluation [16, 17], though we do not know of any existing Bayesian approaches to the policy evaluation problem.

In this paper we develop a flexible Bayesian method to evaluate decision algorithms in the presence of unmeasured confounding. In contrast to most past work on sensitivity analysis, our approach conveniently accommodates latent confounding that varies with the observed covariates. This flexibility, coupled with a full Bayesian specification, helps to automatically exclude values of the latent variables that are not supported by the data and model of confounding.

To illustrate our method, we consider judicial bail decisions, starting with a detailed dataset of over 165,000 judgments in a large,

urban jurisdiction. Based on this information, we create realistic synthetic datasets for which potential outcomes and selection mechanisms are, by construction, fully known. We apply our method to these synthetic datasets, selectively masking potential outcomes and the data-generation process to mimic real-world applications. Under a variety of conditions, we find that our approach accurately recovers true policy effects despite the presence of unmeasured confounding. In these simulations, our approach outperforms classical, non-Bayesian sensitivity analysis techniques that have recently been adapted to the algorithmic decision-making context [16, 21]. Further, when applied to the special case of estimating average treatment effects, our general method performs on par with existing, state-of-the-art approaches that are tailored to that specific problem.

2 METHODOLOGY

A common goal of causal inference is to estimate the average effect of a binary treatment $T \in \{0, 1\}$ on a response $Y \in \mathbb{R}$. That is, one often seeks to estimate $\mathbb{E}[Y(1) - Y(0)]$, where $Y(0)$ and $Y(1)$ denote the potential outcomes under the two possible treatment values. Here we consider a generalization of this problem that arises in many applied settings. Namely, given a policy $\pi : \mathbb{R}^m \mapsto \{0, 1\}$ that assigns treatment based on individual characteristics $X \in \mathbb{R}^m$, we seek to estimate the average response $V^\pi = \mathbb{E}[Y(\pi(X))]$. In our judicial bail application (described in detail in following sections), π is an algorithmic rule that determines which defendants are released on their own recognizance (RoR) and which are required to pay bail, $Y \in \{0, 1\}$ indicates whether a defendant appears at trial, X is a vector of observed covariates, and V^π is the expected proportion of defendants who fail to appear in court when the rule is followed.

2.1 Policy evaluation without unmeasured confounding

To motivate the standard approach for estimating V^π, we decompose it into the sum of two terms:

$$\begin{aligned} V^\pi =& \mathbb{E}[Y(T) \mid \pi(X) = T] \cdot \Pr[\pi(X) = T] \\ &+ \mathbb{E}[Y(1-T) \mid \pi(X) \neq T] \cdot \Pr[\pi(X) \neq T]. \end{aligned}$$

The proposed and observed policies are the same for the first term but differ for the second term. Thus estimating the first term is straightforward, since we directly observe the outcomes for this group, $Y^{\text{obs}} = Y(T)$. The key statistical difficulty is estimating the second term, for which we must impute the missing potential outcomes, $Y^{\text{mis}} = Y(1-T)$. That is, in these cases we must estimate what would have happened had the proposed action been followed, rather than the observed action. This challenge is known as the fundamental problem of causal inference [14].

To estimate V^π, it is common to start with a sample of historical data $\{(x_i, t_i, y_i(t_i))\}_{i=1}^n$ on individual covariates, treatment decisions, and observed outcomes. We can then estimate V^π via:

$$\hat{V}^\pi = \frac{1}{n}\left[\sum_{\pi(x_i)=t_i} y_i^{\text{obs}} + \sum_{\pi(x_i)\neq t_i} \hat{y}_i^{\text{mis}}\right], \tag{1}$$

where y_i^{obs} is the observed outcome for individual i when the proposed and observed policies agree. A simple approach to estimating the unobserved potential outcomes $\hat{y}_i^{\text{mis}}$ in the second term is to directly model the outcomes conditional on treatment and observed covariates—a strategy sometimes referred to as *response surface modeling* [13].

This direct modeling approach implicitly assumes that the treatment is *ignorable* given the observed covariates (i.e., that there is no unmeasured confounding). Formally, ignorability means that

$$Y(0), Y(1) \perp\!\!\!\perp T \mid X. \tag{2}$$

In other words, conditional on the observed covariates, those who receive treatment are similar to those who do not. In many realistic situations, ignorability is a strong assumption. The main contribution of this paper is a Bayesian approach to assessing the sensitivity of estimated policy outcomes V^π to violations of ignorability, which we present next.

2.2 Policy evaluation with unmeasured confounding

When ignorability does not hold, the resulting estimates of V^π can be strongly biased. To address this issue, the sensitivity literature typically starts by assuming that there is an unmeasured confounder $U \in \{0, 1\}$ (or, more generally, $U \in \mathbb{R}$) for each individual such that ignorability holds given both X and U:

$$Y(0), Y(1) \perp\!\!\!\perp T \mid X, U. \tag{3}$$

One then generally imposes additional structural assumptions on the form of U and its relationship to decisions and outcomes. We follow this basic template to obtain sensitivity estimates for policy outcomes.

At a high level, we model the observed data $\{(x_i, t_i, y_i(t_i))\}_{i=1}^n$ as draws from parametric distributions depending on the measured covariates x_i and the unmeasured, latent covariates u_i:

$$\begin{aligned} y_i(0) &\sim f(x_i, u_i; \alpha) \\ y_i(1) &\sim g(x_i, u_i; \beta) \\ t_i &\sim h(x_i, u_i; \gamma) \\ y_i &= y_i(t_i) \end{aligned}$$

where α, β, γ, and u_i are latent parameters with weakly informative priors. The inferred joint posterior distribution on these parameters then yields estimates of $y_i(0)$ and $y_i(1)$ for each individual i. Finally, the posterior estimates of the potential outcomes yield a posterior estimate of V^π via Eq. (1) that accounts for unmeasured confounding.

Our strategy, while conceptually straightforward, differs in three important ways from classical sensitivity methods. First, we directly model the treatment and potential outcomes, allowing for flexibility in the functional forms of f, g and h. Second, our Bayesian approach automatically excludes values of the latent variables that are not supported by the data [19]. In contrast, previous methods generally require more extensive parameter tuning to obtain reasonable estimates. Finally, our focus is on the more general problem of estimating policy outcomes rather than the average treatment effects considered in past work.

We now describe the specific forms of f, g, and h that we use throughout our analysis. First, we reduce the dimensionality of the observed covariate vectors x down to three key quantities:

$\mu_0(x) = \mathbb{E}[Y(0) \mid X = x]$, the conditional average response if $T = 0$; $\mu_1(x) = \mathbb{E}[Y(1) \mid X = x]$, the conditional average response if $T = 1$; and $e(x) = \Pr(T = 1 \mid X = x)$, the propensity score. The first two quantities, $\mu_0(x)$ and $\mu_1(x)$, depend on potentially unobserved outcomes $Y(0)$ and $Y(1)$, and so typically cannot be perfectly estimated from available data—indeed, accurately estimating these terms is our ultimate goal. Thus, as a first step, we generate approximations $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$ by assuming ignorability and fitting standard prediction models, like regularized regression, to subsets of the data with $T = 0$ and $T = 1$, respectively.

To support complex relationships between the predictors and outcomes, we divide the data into K approximately equally sized groups, ranking and binning by the estimated outcome $\hat{\mu}_0$. As Franks et al. [11] show, this flexible model form also plays an important theoretical role, helping to avoid issues of identifiability that are inherently at odds with the exercise of sensitivity analysis. Denoting the group membership of observation i by $k[i] \in \{1, 2, \ldots, K\}$, and a Bernoulli distribution with mean p as $\mathrm{B}(p)$, we model the observed data as follows:

$$
\begin{aligned}
y_i(0) &\sim \mathrm{B}\left(\mathrm{logit}^{-1}\left(\alpha_{0,k[i]} + \alpha_{\hat{\mu}_0,k[i]}\hat{\mu}_0(x_i) + \alpha_{u,k[i]}u_i\right)\right) \\
y_i(1) &\sim \mathrm{B}\left(\mathrm{logit}^{-1}\left(\beta_{0,k[i]} + \beta_{\hat{\mu}_1,k[i]}\hat{\mu}_1(x_i) + \beta_{u,k[i]}u_i\right)\right) \\
t_i &\sim \mathrm{B}\left(\mathrm{logit}^{-1}(\gamma_{0,k[i]} + \gamma_{\hat{e},k[i]}\hat{e}(x_i) + \gamma_{u,k[i]}u_i)\right) \\
y_i &= y_i(t_i).
\end{aligned}
$$

In each of the first three equations above, variables are modeled as draws from a Bernoulli distribution whose mean depends on both the reduced covariates—$\hat{\mu}_0(x)$, $\hat{\mu}_1(x)$, and $\hat{e}(x)$—and the unmeasured confounder u.[1] In the judicial context, for example, one can imagine that u_i corresponds to unobserved risk of failing to appear, with judges more likely to demand bail from riskier defendants.

Finally, to complete the Bayesian model specification, we must describe the prior distribution on the parameters (additional detail is provided in the Appendix). On each of the unmeasured confounders u_i, we set a $\mathrm{N}(0, 1)$ prior. On the coefficients α, β, and γ we use a random-walk prior; intuitively, these priors ensure that adjacent groups have similar coefficient values, mitigating the dependence of results on the exact number of groups K.

By definition, it is impossible to infer the degree of unmeasured confounding from the observed data alone. As such, most methods for sensitivity analysis involve one or more parameters that allow one to *explicitly* set the degree of confounding, which may, for example, be informed by domain knowledge. However, in our case, the necessary side information enters *implicitly*, from the structural form of our model of confounding, especially the choice of K discussed above. By positing smoothness and accordingly restricting the number of groups—which is reasonable in many applications—one can leverage the functional form of the model to avoid having to explicitly set the degree of confounding. This general approach is a valuable complement to the traditional perspective, but there is no free lunch: if the model of confounding is unreasonable for a particular application, the resulting sensitivity bounds will suffer. Below we discuss a strategy for appropriately setting the model parameters via simulation.

[1] Although we describe our method for binary potential outcomes, the general approach can accommodate real-valued outcomes as well.

3 AN APPLICATION TO JUDICIAL DECISION MAKING

To demonstrate our general approach to offline policy evaluation, we now consider in detail the case of algorithms designed to aid judicial decisions [12, 16, 18]. In the U.S. court system, pretrial release determinations are among the most common and consequential decisions for criminal defendants. After arrest, a defendant is usually arraigned in court, where a prosecutor presents a written list of charges. If the case proceeds to trial, a judge must decide whether the defendant should be released on his or her own recognizance (RoR) or subject to money bail, where release is conditional on providing collateral meant to ensure appearance at trial. Defendants who are not RoR'd and who cannot post bail themselves may await trial in jail or pay a bail bondsman to post bail on their behalf. Judges must therefore balance the burdens of bail on a defendant and society against the risk that the defendant may fail to appear (FTA) for trial.[2]

Here we consider algorithmic policies for assisting these judicial decisions, recommending either RoR or bail based on recorded characteristics of a defendant and case. The policy evaluation problem is to estimate, based only on historical data, the proportion of defendants who would fail to appear if algorithmic recommendations were followed. This is statistically challenging because one does not always observe what would have occurred had the algorithmic policy been followed. In particular, if the policy recommends releasing a defendant who was in reality detained, or recommends detaining a defendant who was in reality released, the relevant counterfactual is not observed. Further, since judges may—and likely do—base their decisions in part on information that is not recorded in the data, direct outcome models ignoring unmeasured confounding will likely be biased. We thus allow for a real-valued unobserved covariate u that affects both a judge's decision (RoR or bail) and also the outcome (FTA or not) conditional on that decision. Our goal is to assess the sensitivity of flight risk estimates to such unmeasured confounding.

3.1 Policy construction

Our analysis is based on 165,000 adult cases involving nonviolent offenses charged by a large urban prosecutor's office and arraigned in criminal court between 2010 and 2015. These cases do not include instances where defendants accepted a plea deal at arraignment, where no pretrial release decision is necessary. For each case, we have 49 features describing characteristics of the current charges (e.g., theft, gun-related), and 15 features describing characteristics of the defendant (e.g., gender, age, prior arrests). We also observe whether the defendant was RoR'd, and whether the defendant failed to appear at any of his or her subsequent court dates. Applying notation introduced in the previous section, x_i refers to a vector of all observed characteristics of the i-th defendant, $t_i = 1$ if bail was set, and $y_i = 1$ if the defendant failed to appear at court.

To carry out our analysis, we first randomly select 10,000 cases from the full dataset which we set aside as our final test data. The remaining data are split into two folds of approximately equal size.

[2] In many jurisdictions, judges may also consider the risk that a defendant will commit a new crime if released when deciding whether or not to set bail, but not in the jurisdiction we consider.

We begin by constructing a family of algorithmic decision rules. On the first training fold, we fit an L^1-regularized (lasso) logistic regression model to estimate the probability of FTA given release, $\hat{\mu}_0^{\pi}(x_i)$. That is, we fit a logistic regression with the left-hand side indicating whether a defendant failed to appear at any court dates, and the right-hand side comprised of all available covariates for the subset of defendants that the judge released. We use the superscript π to indicate that these estimates are computed on the first fold of data, and are used exclusively to define policies, not to evaluate them. With these estimates in hand, we construct a family of policies $\{\pi_s\}$, indexed by a risk threshold s for releasing individuals:

$$\pi_s(x_i) = \begin{cases} 1 & \text{if } \hat{\mu}_0^{\pi}(x_i) > s \\ 0 & \text{if } \hat{\mu}_0^{\pi}(x_i) \leq s. \end{cases}$$

These policies, which are based on a ranking of defendants by risk, are similar to pretrial algorithms used in practice [7].

Given the family of policies $\{\pi_s\}$ defined above, we next estimate the proportion of defendants who would fail to appear under each policy, accounting for unmeasured confounding. To do so, on the second training fold, we fit three L^1-regularized logistic regression models to estimate each individual's likelihood of failing to appear if RoR'd or required to pay bail—$\hat{\mu}_0(x_i)$ and $\hat{\mu}_1(x_i)$, respectively—as well as each individual's likelihood of having bail set, $\hat{e}(x_i)$.

Finally, on the test fold consisting of 10,000 cases, we fit our Bayesian sensitivity model using the estimated quantities $\hat{\mu}_0(x_i)$, $\hat{\mu}_1(x_i)$, and $\hat{e}(x_i)$. But, to do so, we must first select appropriate parameters for our model of confounding, which we describe next.

3.2 Calibrating the model via simulation

As mentioned above, information about unmeasured confounding enters through the structural form of the model. In particular, the amount of unmeasured confounding is related to the number of observations per bin, n/K, with fewer observations per bin corresponding to more unmeasured confounding. We now describe how to calibrate our model of unmeasured confounding using synthetic datasets, before applying our method to real data.

We begin by creating a synthetic version of our original judicial dataset of judicial decisions where *both* potential outcomes for each defendant—$y_i(0)$ and $y_i(1)$—are known. We do this by first estimating $\hat{\mu}_0(x_i)$, $\hat{\mu}_1(x_i)$, and $\hat{e}(x_i)$ using L^1-regularized logistic regression, as above. Then, for each individual in the original dataset, we create synthetic potential outcomes and treatment assignments via independent Bernoulli draws based on these estimated probabilities. We denote the resulting synthetic (uncensored) dataset by $\Omega = \{(x_i, t_i', y_i'(0), y_i'(1))\}$. Because both potential outcomes are known, we can exactly calculate V^{π} for any policy π applied to Ω. Moreover, our synthetic dataset satisfies ignorability by construction.

We next censor Ω in two ways. First, we restrict the observed covariates to a subset x_i'. Specifically, we consider three sets of restricted covariates: age; age and gender; and age, gender, and prior number of missed court appearances. Second, for each individual we remove the potential outcome for the action not taken, keeping only $y_i'(t_i')$. Thus, for each of our three choices of x_i', we have a dataset of the form $\Omega' = \{(x_i', t_i', y_i'(t_i'))\}$. The covariates not included in x_i' correspond to unmeasured confounding. Starting from these

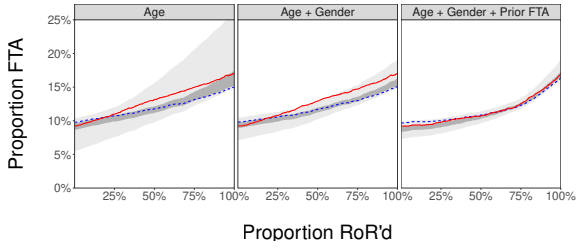

Figure 1: ***Sensitivity estimates on three synthetic datasets. In each panel, datasets are restricted to contain only a subset of covariates: age (left); age and gender (center); age, gender, and prior FTAs (right). The blue lines show estimates based on direct outcome models ignoring unmeasured confounding, and the gray bands represent 50% and 95% credible intervals based on our sensitivity analysis. The red lines show the true policy outcomes.***

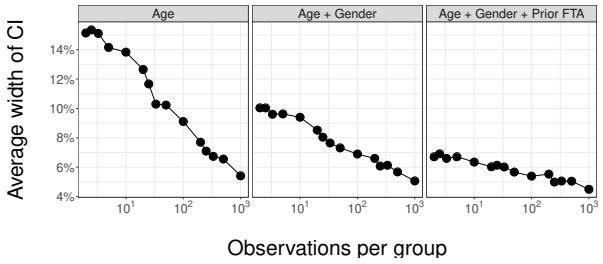

Figure 2: ***The average width of 95% credible intervals for a range of n/K observations per group.***

three synthetic datasets, we carry out the policy construction and sensitivity analysis procedure described above.

The results of our simulation study are shown in Figure 1, with each panel corresponding to a different choice of x_i' and thus a different degree of unmeasured confounding. In all cases, we set $n/K = 1000$ (i.e., 1,000 observations per bin). The blue lines show estimates based on direct outcome modeling, ignoring unmeasured confounding, and the gray bands represent 50% and 95% credible intervals based on our sensitivity analysis. Importantly, because these results are derived from synthetic datasets, we can also compute the true policy outcomes, which are indicated in the plot by the red lines.

Across the three levels of unmeasured confounding that we consider, our sensitivity bands cover the ground truth line. Moreover, the bands accurately reflect the true level of confounding, with wider bands when x_i' consists only of age (i.e., there is extreme confounding), and narrower bands when x_i' is comprised of age, gender, and prior number of missed court appearances (the confounding is less severe). Thus, based on this simulation, setting $n/K = 1000$ appears reasonable to account for the true level of unmeasured confounding in the synthetic datasets.

To illustrate the effect of our parameter choice on the results, we measure the average width of the 95% credible intervals in our synthetic datasets over a wide range of n/K (i.e., observations per bin). The results are presented in Figure 2, which is based on a

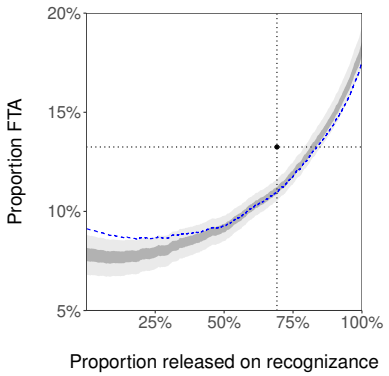

Figure 3: ***Sensitivity estimates for flight risk under a family of policies π_s. The blue line denotes estimates assuming no unmeasured confounding. The gray bands indicate 95% and 50% credible intervals under our model of unmeasured confounding.***

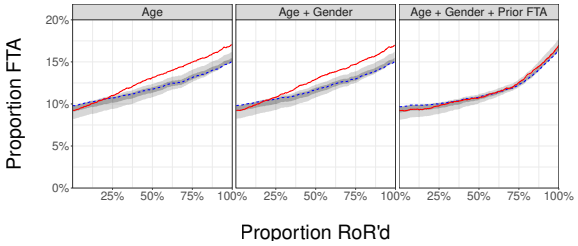

Figure 4: ***Sensitivity estimates based on the method of Jung et al. [16], applied to three synthetic datasets. The gray bands indicate minimum and maximum policy estimates for two parameter regimes where the unobserved confounder can both double the odds of bail and of failing to appear at trial (dark bands), or triple these odds (light bands). The blue lines show estimates from direct outcome models ignoring unmeasured confounding, and the red lines indicate ground truth outcomes.***

dataset of 1000 observations. As the number of observations per bin increases—as a result of lowering the number of groups K—the credible interval widths decrease. Importantly, the credible intervals for all the settings we consider still cover the ground truth in the synthetic datasets, as in Figure 1.

3.3 Results on the observed data

Based on the above calibration exercise, we fit our Bayesian sensitivity model on the real data using $n/K = 1000$ observations per bin (i.e, with $K = 10$ bins on the held-out set of 10,000 data points). We note that we found qualitatively similar results when we repeated our analysis for $K = 5$ and $K = 100$ bins.

The results of our analysis are plotted in Figure 3. The blue dashed line shows, for each policy π_s, the proportion RoR'd and the estimated proportion that FTA under the policy, $\hat{V}^{\pi_s}$, estimated assuming no unobserved confounding. For reference, the black point shows the status quo: judges in our dataset release 69% of defendants, resulting in an overall FTA rate of 13%. The light and dark gray bands show, respectively, the 95% and 50% credible intervals that result from the sensitivity analysis procedure under our model of unobserved confounding.

Figure 3 illustrates three key points. First, for policies that RoR almost all defendants (toward the right-hand side of the plot), the blue line lies below our sensitivity bands, in line with expectations. If there is unmeasured confounding, we would expect those who were in reality detained to be riskier than they seem from the observed data alone; as a result, the direct outcome model underestimates the proportion of defendants that would fail to appear if all (or almost all) such defendants are RoR'd. Conversely, for policies that recommend bail for almost all defendants (toward the left-hand side of the plot), the blue line lies above the sensitivity bands, as we would expect, because those who were in reality released are likely less risky than they appear from the observed data alone. Second, as the policies move further from the status quo, toward the left- and right-hand extremes of the plot, the sensitivity bands grow in width, indicating greater uncertainty. This pattern reflects the fact that the data provide less direct evidence of flight risk as policies diverge from observations, heightening the potential impact of unmeasured confounding. Finally, even after accounting for unmeasured confounding, there are algorithmic policies that do substantially better than the status quo, in the sense that they RoR more defendants and simultaneously achieve a lower overall FTA rate.

4 COMPARISON WITH ALTERNATIVE METHODS

Sensitivity analysis for offline policy evaluation is relatively new [16, 17]. There are, however, several methods for assessing the sensitivity of average treatment effects to unmeasured confounding, which is a specific case of policy evaluation. Here we compare our approach to Jung et al.'s sensitivity method for offline policy evaluation, as well as to two recent Bayesian sensitivity methods designed for average treatment effects [9, 19]. For space, we omit the details of these methods, but note that we implement the two sensitivity analysis procedures exactly as they were originally described.

Figure 4 shows the results of the Jung et al. procedure applied to the three synthetic datasets described above. In particular, we compute the minimum and maximum policy estimates obtained by sweeping over two parameter regimes suggested in their paper. In the first regime (dark bands), we allow the unobserved confounder to double the odds of bail and the odds of failing to appear at trial, both if RoR'd or required to pay bail. We also consider a more extreme situation (lighter bands), where the unobserved confounding can increase the odds of being detained up to three times, and also triple the odds a defendant fails to appear. In each scenario, the red lines indicate the true outcomes of each policy, computed from the uncensored synthetic dataset, the blue lines show estimates based on direct outcome models ignoring unmeasured confounding, and the gray bands indicate the minimum and maximum values of each policy over all parameter settings in the corresponding regime.

In contrast to our own sensitivity analysis results, the sensitivity bands in Figure 4 often fail to capture the ground truth policy estimates. Further, and more importantly, the sensitivity bands do not

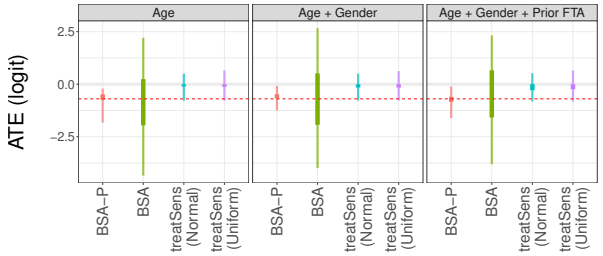

Figure 5: ***Estimates of average treatment effects for three synthetic datasets with differing levels of unmeasured confounding, as estimated by various methods: our own approach, labeled BSA-P; the method of McCandless and Gustafson [19], labeled BSA; and the method of Dorie et al. [9], labeled TreatSens, for two different prior distributions. The thick and thin lines represent 50% and 95% credible intervals, and the red horizontal line marks the true effect.***

appropriately adapt to the differing levels of unmeasured confounding across datasets, as indicated by their relatively constant width across settings. As a result, one must manually tune the sensitivity parameters for each dataset to achieve satisfactory performance. While not impossible—and indeed such calibration is the norm in classical sensitivity methods—the need for such manual adjustment is a significant limitation of non-Bayesian approaches to sensitivity analysis.

Since policy evaluation is a generalization of estimating average treatment effects, we next compare our approach to methods designed for that specific problem. In our judicial application, the average treatment effect is the difference in the proportion of defendants who fail to appear at court when all are required to pay bail versus all being RoR'd. We specifically compare our approach to two recently proposed methods for Bayesian sensitivity analysis of average treatment effects: the method of McCandless and Gustafson [19], which we refer to as BSA, and the TreatSens method of Dorie et al. [9]. Figure 5 shows the results of estimating average treatment effects on our three synthetic datasets, where the true answer is indicated by the dashed red line, and our approach is labeled BSA-P. For TreatSens, since we can specify different priors on the unobserved u, we compute results with a standard normal prior—as used in our own method—and with uniform priors, as suggested by Dorie et al. The thick and thin lines show the 50% and 95% credible intervals. In all three synthetic datasets, our approach is competitive with, and arguably even better than, the two methods we compare to. In particular, whereas the true answer is at the periphery of the 95% credible intervals generated by TreatSens, the ground truth lies near the posterior median of our approach. Further, our credible intervals are substantially narrower than those resulting from BSA, indicating that our method simultaneously achieves accuracy and precision.

5 DISCUSSION

We have addressed the problem of offline policy evaluation by coupling classical statistical ideas on sensitivity analysis with modern methods of machine learning and large-scale Bayesian inference. The result is a conceptually simple yet powerful technique for evaluating algorithmic decision rules.

By definition, it is impossible to precisely quantify *unmeasured* confounding, and so all methods of sensitivity analysis require assumptions that are inherently untestable [11]. Traditional methods handle this situation by requiring practitioners to specify parameters describing the structure and scale of the assumed confounding, informed, for example, by domain expertise. In our case, the necessary side information enters through the assumed form of the data-generating process. By adopting an expressive model form that we calibrate through simulation, our approach balances the need to provide at least some information about the structure of the potential confounding with the impossibility of specifying it exactly. This middle ground appears to work well in practice, but it is useful to remember the conceptual underpinnings of our strategy when applying it to new domains.

Over the last two decades, sophisticated methods of machine learning have emerged and gained widespread adoption. More recently, these methods have been applied to traditional problems of causal inference and their modern incarnations, like offline policy evaluation. Prediction and causal inference are two sides of the same coin, but the links between the two are still under-developed. Here we have bridged one such gap by porting ideas from classical sensitivity analysis to algorithmic decision making. Looking forward, we hope that sensitivity analysis is more tightly integrated into machine-learning pipelines and, more generally, that our work spurs further connections between methods of causal inference and prediction.

REFERENCES

[1] Susan Athey and Stefan Wager. 2017. Efficient policy learning. *arXiv preprint arXiv:1702.02896* (2017).
[2] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016).
[3] Richard Berk. 2012. *Criminal justice forecasts of risk: a machine learning approach.* Springer Science & Business Media.
[4] Nicole Bohme Carnegie, Masataka Harada, and Jennifer L Hill. 2016. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness* 9, 3 (2016), 395–420.
[5] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency.* 134–148.
[6] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *https://arxiv.org/abs/1808.00023* (2018). arXiv:1808.00023
[7] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 797–806.
[8] Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. 1959. Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Nat. Cancer Inst* 22 (1959), 173–203.
[9] Vincent Dorie, Masataka Harada, Nicole Bohme Carnegie, and Jennifer Hill. 2016. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine* 35, 20 (2016), 3453–3470.
[10] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly Robust Policy Evaluation and Learning. *ICML* (2011). https://doi.org/10.1214/14-STS500
[11] Alexander M. Franks, Alexander D'Amour, and Avi Feller. 2019. Flexible Sensitivity Analysis for Observational Studies Without Observable Implications. *J. Amer. Statist. Assoc.* 0, 0 (2019), 1–33. https://doi.org/10.1080/01621459.2019.1604369 arXiv:https://doi.org/10.1080/01621459.2019.1604369
[12] Sharad Goel, Ravi Shroff, Jennifer L Skeem, and Christopher Slobogin. 2018. The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment. (2018). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3306723.

[13] Jennifer L Hill. 2012. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* (2012).
[14] Paul W Holland. 1986. Statistics and Causal Inference. *J. Amer. Statist. Assoc.* 81, 396 (1986), 945–960.
[15] Guildo W Imbens. 2003. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* 93, 2 (2003), 126–132.
[16] J. Jung, C. Concannon, R. Shroff, S. Goel, and D. G. Goldstein. 2017. Simple rules for complex decisions. *ArXiv e-prints* (Feb. 2017). arXiv:stat.AP/1702.04690
[17] Nathan Kallus and Angela Zhou. 2018. Confounding-Robust Policy Improvement. In *Advances in Neural Information Processing Systems*.
[18] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The Quarterly Journal of Economics* 133, 1 (2017), 237–293.
[19] Lawrence C McCandless and Paul Gustafson. 2017. A comparison of Bayesian and Monte Carlo sensitivity analysis for unmeasured confounding. *Statistics in Medicine* (2017).
[20] Lawrence C McCandless, Paul Gustafson, and Adrian Levy. 2007. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in medicine* 26, 11 (2007), 2331–2347.
[21] Paul R Rosenbaum and Donald B Rubin. 1983. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)* (1983), 212–218.
[22] Ravi Shroff. 2017. Predictive Analytics for City Agencies: Lessons from Children's Services. *Big Data* 5, 3 (2017), 189–196.
[23] Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. 2012. A robust method for estimating optimal treatment regimes. *Biometrics* 68, 4 (2012), 1010–1018.
[24] Yichi Zhang, Eric B Laber, Anastasios Tsiatis, and Marie Davidian. 2015. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* 71, 4 (2015), 895–904.

APPENDIX

In the main text, we described the likelihood function for our Bayesian model of unmeasured confounding. Here we complete the model specification by describing the prior distribution on the parameters.

As mentioned in the main text, we set a $\mathrm{N}(0,1)$ prior on each of the unmeasured confounders u_i. On the coefficients α, β, and γ we use a random-walk prior. Formally, the random-walk prior on α_0 is given by

$$\begin{aligned}\alpha_{0,1} &\sim \mathrm{N}(0,1)\\ \alpha_{0,j} &\sim \mathrm{N}(\alpha_{0,j-1}, \tau^2_{\alpha_0}) \quad \text{for } j \in \{2,3,\ldots,K\}\\ \tau_{\alpha_0} &\sim \mathrm{N}_+(0,\sigma^2_\tau),\end{aligned}$$

where $\mathrm{N}_+(0,\sigma^2_\tau)$ indicates the half-normal distribution with standard deviation σ_τ. We analogously set priors for the parameters $\alpha_{\hat{\mu}_0}$, β_0, $\beta_{\hat{\mu}_1}$, γ_0, and $\gamma_{\hat{e}}$.

For the coefficients on the unobserved confounders u_i, we set random-walk priors with an additional constraint to ensure positivity, so that $\Pr(T=1)$, $\Pr(Y(0)=1)$, and $\Pr(Y(1)=1)$ all increase with u_i. This constraint, while not strictly necessary, facilitates estimation of the posterior distribution. Formally, for α_u we have

$$\begin{aligned}\alpha_{u,1} &\sim \mathrm{N}_+(0,1)\\ \alpha_{u,j} &\sim \mathrm{N}_+(\alpha_{u,j-1}, \tau^2_{\alpha_u}) \quad \text{for } j \in \{2,3,\ldots,K\}\\ \tau_{\alpha_u} &\sim \mathrm{N}_+(0,\sigma^2_\tau).\end{aligned}$$

We similarly set sign-constrained random-walk priors on β_u and γ_u. For the main results in this paper, we set values of $\sigma_\tau = 1$ and $K = 10$. However, the results are not substantially affected by the choice of K or the prior distribution (parameterized by σ_τ).