Published in final edited form as:

Biom J. 2018 July ; 60(4): 721–733. doi:10.1002/bimj.201700305.

Bayesian propensity scores for high-dimensional causal inference: a comparison of drug-eluting to bare-metal coronary stents

Jacob V Spertus¹ and Sharon-Lise T Normand^{1,2,*}

¹: Department of Health Care Policy, Harvard Medical School, Boston

²: Department of Biostatistics, Harvard TH Chan School of Public Health, Boston

Abstract

High-dimensional data provides many potential confounders that may bolster the plausibility of the ignorability assumption in causal inference problems. Propensity score methods are powerful causal inference tools, which are popular in health care research and are particularly useful for high-dimensional data. Recent interest has surrounded a Bayesian treatment of propensity scores in order to flexibly model the treatment assignment mechanism and summarize posterior quantities while incorporating variance from the treatment model. We discuss methods for Bayesian propensity score analysis of binary treatments, focusing on modern methods for high-dimensional Bayesian regression and the propagation of uncertainty. We introduce a novel and simple estimator for the average treatment effect that capitalizes on conjugacy of the beta and binomial distributions. Through simulations, we show the utility of horseshoe priors and Bayesian additive regression trees paired with our new estimator, while demonstrating the importance of including variance from the treatment regression model. An application to cardiac stent data with almost 500 confounders and 9000 patients illustrates approaches and facilitates comparison with existing alternatives. As measured by a falsifiability endpoint, we improved confounder adjustment compared with past observational research of the same problem.

Keywords

Bayesian methods; propensity score weighting; regularization; Bayesian additive regression trees; coronary stent

1 Introduction

Comparative effectiveness research examines which health care interventions work best for which patients. These questions are typically causal in nature and frequently informed by randomized clinical trials (RCTs). When experimental evidence is unavailable, causal inferences may be drawn from observational data such as information found in health care

*Corresponding Author: sharon@hcp.med.harvard.edu, 617-432-3260.

Conflict of Interest Statement

The authors have declared no conflict of interest.

databases. Propensity score approaches that condition on treatment assignment probabilities are often used to adjust for confounding and attain unbiased causal effects provided key assumptions are met. They have a central role in applied causal inference for inferring effectiveness of interventions in usual care settings or in regulatory settings for clearance of medical devices (U.S. Food and Drug Administration, 2013). One key feature to the validity of findings is that propensity score approaches build treatment models without reference to outcome information in analogy to randomized trials, an outcome free design. This feature is vital for high-dimensional causal inference problems where many potential confounders are available (Rubin, 2008)(Austin and Stewart, 2015)(Schneeweiss et al., 2009).

Recent work has attempted to bring propensity scores into the domain of Bayesian inference. Although difficult to reconcile with traditionally frequentist propensity scores, Bayesian methods offer their own suite of tools for model building, provide intuitive ways to present posterior summaries of parameters, and, particularly important for high-dimensional causal inference, propagate uncertainty from all stages of model fitting (Kaplan and Chen, 2012)(Zigler and Dominici, 2014)(Saarela et al., 2015)(Graham et al, 2016). These features make Bayesian propensity scores an appealing area for development.

Our work is motivated by a comparison of the effectiveness of two treatment alternatives for percutaneous coronary intervention (PCI) in a cohort of patients treated in Massachusetts non-federal hospitals. These interventions are standard care for clearing blockages in the blood vessels supplying the heart and frequently include the installation of a coronary stent to keep the treated artery clear and supported. We focus on two classes of coronary stents available in the U.S. market: bare-metal (BMS) which were first approved in the U.S. in the 1990s and drug-eluting (DES) approved in the U.S. in 2003. Many randomized clinical trials, meta-analyses, and observational studies have compared the effectiveness of DESs and BMSs for patients with a range of conditions (Venkitachalam et al., 2011)(Mauri et al., 2008)(Bonaa et al., 2016). Our main data source is the Mass-DAC clinical registry, a state mandated database harvesting clinical information for all PCIs performed in adults (age 18 years) in all non-federal Massachusetts' hospitals annually. We sought to estimate the causal effect of DES compared to BMS on all-cause mortality and a composite adverse event endpoint within 1 year of the initial implant. Our data, examining stents implanted in 2011, includes about 9,000 patients undergoing coronary stenting and approximately 500 confounders. A small subset of these confounders and outcomes appears in table 1. Generally, BMS patients are sicker and older than DES patients, confounding a naive estimate of causal effect.

We propose a method of Bayesian propensity score analysis for causal inference, with emphasis on using high-dimensional data to meet the ignorability assumption. Our method estimates a difference in outcome proportions in a weighted pseudo-population, sharing links to frequentist inverse probability weighting (IPW) as well as the weighted likelihood approach of Saarela et al. (2015). By focusing on a Bayesian framework, we allow for intuitive and flexible approaches to model building with many variables while propagating uncertainty in the treatment model. We do not offer new theory on Bayesian propensity scores, referring the interested reader to the extensive discussion in Saarela et al. (2015) and its accompanying critiques, as well as Zigler (2016). Our interest is on the potential utility of

a two-step Bayesian propensity score framework in high-dimensional causal inference problems. We assess our approach on the basis of finite sample frequentist characteristics such as consistency, efficiency, and coverage through simulation studies.

We next state the goals and assumptions of observational causal inference in section 2. Section 3 briefly reviews the literature on Bayesian propensity scores and high-dimensional causal inference. In section 4 we provide details on treatment model regression and propose a two-step propensity score weighting procedure using Bayesian methods. In section 5 we characterize finite sample performance with simulation studies. We return to the coronary stent problem in section 6. Throughout we assume a binary treatment, referring to one arm as treatment and the other arm generically as control. We label the “treatment model” as the model characterizing the relationship between confounders and treatment, and the “outcome model” as the model characterizing the relationship between treatment and outcome, adjusting for confounders.

2 Assumptions for Causal Inference

We make three assumptions necessary for causal inference: positivity, stable unit treatment value assignment (SUTVA), and ignorability. SUTVA encodes the assumptions that observations do not interfere with one another and that the nature of treatment does not vary across individuals. Positivity refers to the assumption that each subject has a chance for assignment to either treatment group. It is difficult to test for positivity violations in high-dimensions, as each subject is typically uniquely defined by some combination of their covariates. We do assume positivity is met in our data and refer the reader to our references for more information (Petersen et al., 2010). Our focus is on meeting the ignorability assumption by exploiting high-dimensional data.

2.1 Ignorability

For a subject i , let X_i denote binary treatment, Y_{1i} and Y_{0i} potential outcomes under treatment and control respectively, and C_i a vector of pre-treatment confounders. We are interested in the treatment effect, $E(Y_{0i}) - E(Y_{1i})$. However we can never observe both potential outcomes for a subjects. We can use observational data to estimate the treatment effect if we are willing to assume that conditional on observed pre-treatment confounders, treatment is independent of potential outcomes across the population (Rosenbaum and Rubin, 1983):

$$(Y_{0i}, Y_{1i}) \perp X_i | C_i.$$

Coupled with positivity, ignorability permits estimation of the average treatment effect (ATE) defined here as the difference in marginal event rates under each treatment across the population:

$$\Delta = E_C \{ E(Y | X = 1, C) - E(Y | X = 0, C) \}.$$

If an important confounder is excluded from $\mathbf{C}$, then the ignorability assumption is not met and our estimate of τ will be biased. Thus in the observational setting, inclusion of important confounders is essential to generating unbiased causal estimates.

3 Causal Inference and Bayesian Methods

3.1 Propensity Scores

For binary treatment the propensity score for subject i is defined as $\pi(C_i) = P(X_i = 1 | C_i)$, the probability of receiving treatment given pre-treatment confounders. In their seminal paper, Rosenbaum and Rubin (1983) define the vector $\boldsymbol{\pi}(\mathbf{C}) = [\pi(C_1) \dots \pi(C_n)]$ as a “balancing score” and show that conditioning on it is sufficient to estimate unbiased causal effects, i.e. $(Y_0, Y_1) \perp X | \boldsymbol{\pi}(\mathbf{C})$. This fact is particularly useful in the high-dimensional setting because we only need to condition on the one-dimensional propensity score vector, $\boldsymbol{\pi}(\mathbf{C})$, rather than the entire matrix of confounders, $\mathbf{C}$ (Rosenbaum and Rubin, 1983). However, in the absence of randomization, the propensity scores must still be estimated from a regression of confounders on treatment and a first step involves reducing the dimension of the confounders via variable selection or by some method of regularization. Focusing on the treatment model only, an outcome free design, is preferable over estimates based on outcome regression adjustment because it reduces the risk of selective inference (Rubin, 2008) (Austin and Stewart, 2015). Because this approach reduces investigator bias and shares important connections to randomized trials, propensity score methods are recommended for use in observational studies by important regulators and decision-makers (U.S. Food and Drug Administration, 2013).

3.2 Bayesian Propensity Scores

Several researchers have justified and implemented propensity scores in a fully Bayesian framework by jointly modeling the outcome and treatment assignment mechanism (Zigler and Dominici, 2014). In addition to adhering to the likelihood principle, an advantage of these approaches is that instrumental variables can be screened out by the use of carefully constructed priors that select variables associated with both treatment and outcome (Zigler and Dominici, 2014). Nonetheless, strong critiques have been raised involving fully Bayesian joint propensity score methods. First, when the outcome model parameters are a-priori independent of the treatment model, the propensity scores play no role in Bayesian inference. Second, if a-priori treatment and outcome parameters are dependent then joint modeling of outcome and treatment change the treatment balancing property of the propensity score. Finally, joint modeling can be very sensitive to model misspecification, leading to poor calibration and coverage (Zigler, 2016)(Zigler and Dominici, 2014)(Saarela et al., 2015)(Graham et al, 2016).

We note that an important alternative to propensity score methods are doubly-robust approaches, pioneered by Robins et al. (1994) Robins with augmented-inverse probability weighting. Recent work has created Bayesian techniques for doubly-robust estimation: Saarela et al. (2016) discuss theoretical connections and hurdles to Bayesian doubly-robust estimation, Graham et al (2016) present an approximately Bayesian method based on the Bayesian bootstrap, and Cefalu et al. (2016) describe an algorithm for model-averaged

doubly-robust estimates. Targeted maximum likelihood is another doubly-robust alternative, though one that has been researched solely in a frequentist framework (van der Laan and Rose, 2011).

3.3 High-dimensional Causal Inference

Regularization for causal inference has garnered much attention, mostly predicated on outcome modeling. Recent highlights in the frequentist literature include the lasso based approaches of Ghosh et al. (2015) and Farrell (2015). For propensity score analysis Ghosh (2011) discussed dimension reduction methods and Schneeweiss et al. (2009) developed a claims data variable selection algorithm. On the Bayesian side, Wang et al. (2012) proposed an algorithm using discrete priors with linked treatment and outcome parameters to select likely confounders. Hahn et al. (2016) used horseshoe priors with a special parameterization to estimate a treatment effect in an outcome regression and Zigler and Dominici (2014) proposed a joint Bayesian propensity score approach with discrete regularizing priors.

Non-parametric tree-based approaches have also shown promise. McCaffrey et al. (2004) developed generalized boosted models for frequentist propensity score analysis, and both Hill (2011) and Hahn et al. (2017) demonstrated the utility of Bayesian additive regression trees as an outcome model for causal inference.

4 Proposed Approaches

Along the lines of Kaplan and Chen (2012) and Saarela et al. (2015), we pursue a Bayesian computational approach where the treatment and outcome model are estimated in separate stages. For the treatment model, we propose Bayesian logistic models regularized with weakly informative or sparsity inducing priors, as well as non-parametric Bayesian additive regression trees (BART). For the outcome model, we sample from conjugate beta posterior distributions whose arguments are based on a weighted number of counts observed in the samples.

Such a method is not widely accepted as formally Bayesian because there is not a joint likelihood for all parameters, so we must assess our methods on the basis of features like bias, efficiency, and coverage (Saarela et al., 2015). Nevertheless, such two-step methods are promising causal inference tools (Zigler, 2016). For one, we maintain the outcome free design feature while allowing the use of Bayesian modeling strategies (Bayesian algorithms, priors, posterior sampling, etc) for the propensity score model. Additionally, we propagate uncertainty from both the propensity score and outcome model which may improve frequentist properties of causal estimators in high-dimensions (Zigler and Dominici, 2014). We also generate the full posterior density for the causal parameter of interest, allowing flexible presentation and interpretation of results (Graham et al, 2016). Kass and Steffey (1989) were perhaps the first to identify the utility of such “approximately Bayesian” procedures, focusing on achieving good inferential properties in hierarchical models.

4.1 Regularization

Regularization is a popular technique in high-dimensional data analysis. Suppose we wish to estimate a logistic regression model to predict a binary treatment X from confounders $\mathbf{C}$ with coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$:

$$X_i \sim \text{Bernoulli} \left\{ \text{logit}^{-1} \left(\beta_0 + \sum_{j=1}^p \beta_j C_{ij} \right) \right\}. \quad (1)$$

If p , the dimension of $\mathbf{C}$, is large, the parameter vector $\boldsymbol{\beta}$ becomes highly variable and the model tends to “over-fit” the data. Regularization trades a bit of bias for a relatively large efficiency gain when estimating $\boldsymbol{\beta}$. This controls overfitting and can reduce the mean squared error of estimates. From a Bayesian perspective, regularization amounts to placing informative prior distributions on parameters, often centered at zero. Commonly used regularizing distributions include normal, double-exponential (Bayesian lasso), and Student-t distributions (Gelman et al., 2014). Choosing to exclude variable j based on subject matter knowledge is equivalent to prior certainty that $\beta_j = 0$.

Specification of priors can be very intuitive: we might reasonably assume that a binary confounder will not have a log-odds effect greater than 5 in magnitude, which is quite large in our motivating example. A so called “weakly informative prior,” the Student-t distribution with 3 degrees of freedom and scale 2.5 places most prior density within $[-5, 5]$ but allows larger coefficients to escape if need be (Gelman et al., 2014).

In sparse high-dimensional settings, ideal regularization priors separate confounders from by aggressively shrinking some coefficients towards zero while allowing true confounders to remain unbiased. The horseshoe prior is a Bayesian regularization tool designed for this purpose (Carvalho et al., 2010). It has both an infinite density near zero and fat tails that do not over-shrink true signals. It can be expressed as a scale mixture of normal distributions:

$$\beta_j \sim \mathcal{N}(0, \lambda_j^2 \tau^2) \text{ where } \lambda_j \sim \text{Cauchy}^+(0, 1) \text{ and } \tau \sim \text{Cauchy}^+(0, 1).$$

The parameter τ acts as a global scale parameter and controls the overall degree of shrinkage; λ_j acts as a local scale parameter for each coefficient, allowing confounder coefficients to be quite large while shrinking noisy coefficients towards zero. Weakly informative half-Cauchy priors on λ_j and τ imply a fully Bayesian specification (Gelman et al., 2014). With default specifications, horseshoe priors outperform the cross-validated lasso in sparse prediction problems and provide fully Bayesian inference (Carvalho et al., 2010).

4.2 Bayesian Additive Regression Trees (BART)

BART is a nonparametric modeling technique that translates decision tree-based ensemble methods, such as random forests, to a Bayesian framework. Such approaches are especially desirable when the form of the model is unknown. Chipman et al. (2010) provide a thorough introduction to the method. Broadly speaking, BART is a sum-of-trees model where prior

distributions are placed over the parameters including tree depth, splitting variables, splitting values, and terminal node (leaf) estimates. As in linear models, the priors allow for regularization, while the Bayesian framework allows for posterior averaging and inference using MCMC. However, like other tree-based methods, BART has a number of advantages over linear models in that it automatically attempts to account for non-linearities and interactions between variables. Through simulations and applications, Chipman et al. (2010) showed BART to be highly competitive against advanced machine learning methods in prediction problems. Hill (2011) further demonstrated the efficacy of BART when used as an outcome model for causal inference. Though Hill's paper dealt with relatively few variables, Chipman et al. (2010) showed that BART scales quite well to high-dimension regression problems. Further details of the algorithm and our implementation are included in the supplementary appendix. A recent approach by Hahn et al. (2017) uses a BART outcome model for causal inference while including a fixed estimate of the propensity score for additional adjustment.

4.3 Propensity Score Weighting and Competitors

When used for a treatment regression, the fitted values from regularized Bayesian logistic models and BART are samples from the propensity score distribution. Weighting is a typical way of using propensity scores to get unbiased ATE estimates. Propensity score weighting creates a pseudo-population of subjects from the original sample in which the treatment effect is unconfounded by covariate imbalances between treatment groups. A weighted estimator is given by:

$$\Delta_{IPW} = \left(\sum_{i=1}^n \frac{X_i}{\pi_i} \right)^{-1} \sum_{i=1}^n \frac{X_i Y_i}{\pi_i} - \left(\sum_{i=1}^n \frac{1 - X_i}{1 - \pi_i} \right)^{-1} \sum_{i=1}^n \frac{(1 - X_i) Y_i}{1 - \pi_i} \quad (2)$$

where for brevity $\pi_i = \pi(C_i)$. (2) generates a weighted pseudo-population by allowing every subject to represent $w_i = \frac{X_i}{\pi_i} + \frac{1 - X_i}{1 - \pi_i}$ subjects. Intuitively treated subjects who “look like” untreated subjects (and vice versa) receive more weight.

Theoretical results and simulations studies show weighting to be superior to matching or regression adjustment in some settings, though matching can sometimes be more stable and intuitive in practice (Austin, 2010)(Lunceford and Davidian, 2004). Weighting is also relatively interpretable and easy to use, garnering an increasingly large user base among health care researchers (Austin and Stewart, 2015). For these reasons and to avoid explicit modeling of the outcome regression, we develop below a weighted estimator with close connections to equation (2).

In terms of inference, multiple proposals exist to derive the variance of equation (2), including sample variances and the empirical sandwich method, as well as the non-parametric bootstrap (Lunceford and Davidian, 2004). Bootstrapping approximates the posterior distribution and incorporates uncertainty from the treatment model, serving as a

bridge between typical propensity score analysis and two-step approaches with Bayesian models (Saarela et al., 2015)(Graham et al, 2016).

4.4 Bayesian Propensity Score Weighting

After estimating the propensity score model, we create an unconfounded pseudo-population by weighting each subject using $w_i = \frac{X_i}{\pi_i} + \frac{1-X_i}{1-\pi_i}$. In this pseudo-population, let $\tilde{Y}(0)$ denote the total number of events in the control group and $\tilde{Y}(1)$ denote the total number of events in the treatment group, with fixed population sizes n_0 and n_1 respectively. We use a binomial likelihood $\tilde{Y}(1)|p_1 \sim \text{Binomial}(n_1, p_1)$ where p_1 is the marginal probability of the event in the treatment group. For the control group we analogously have $\tilde{Y}(0)|p_0 \sim \text{Binomial}(n_0, p_0)$. We are interested in inference on the marginal outcome probabilities p_0 and p_1 , which are equal in expectation to the quantities in equation (2) and thus unconfounded. So using Bayes' rule and suppressing the fixed n_0 and n_1 their posteriors are given by:

$$P(p_0|\tilde{Y}(0)) \propto P(\tilde{Y}(0)|p_0) \cdot P(p_0) \text{ and } P(p_1|\tilde{Y}(1)) \propto P(\tilde{Y}(1)|p_1) \cdot P(p_1). \quad (3)$$

For the priors, we assume independent beta distributions $p_0 \sim \text{Beta}(\alpha_{00}, \alpha_{01})$ and $p_1 \sim \text{Beta}(\alpha_{10}, \alpha_{11})$. The hyperparameters, $\{\alpha_{00}, \alpha_{01}, \alpha_{10}, \alpha_{11}\}$, respectively correspond to prior counts of no events in the control group, events in the control group, no events in the treatment group, and events in the treatment group. We set $\{\alpha_{00}, \alpha_{01}, \alpha_{10}, \alpha_{11}\} = \{1, 1, 1, 1\}$, implying a flat prior distribution over $[0,1]$ for p_0 and p_1 , though informative hyperparameters could be specified if appropriate. The beta priors are conjugate to the binomial likelihood and closed form posteriors for p_0 and p_1 are obtained by augmenting the prior counts with those observed in the data. Thus for a given propensity score $\pi(\mathbf{C})$ and observed data $\mathbf{D} = \{\mathbf{Y}, \mathbf{X}\}$ corresponding to the $n \times 2$ matrix of observed binary outcome and treatment vectors, estimated posterior p_0 and p_1 are given by:

$$p_1|\mathbf{D}, \pi(\mathbf{C}) \sim \text{Beta}(a_1, b_1) \text{ and } p_0|\mathbf{D}, \pi(\mathbf{C}) \sim \text{Beta}(a_0, b_0), \text{ where} \quad (4)$$

$$\begin{aligned} a_1 &= \alpha_{11} + \gamma_1 \left(\sum_{i=1}^n \frac{X_i Y_i}{\hat{\pi}_i} \right); & a_0 &= \alpha_{00} + \gamma_0 \left(\sum_{i=1}^n \frac{(1-X_i) Y_i}{1-\hat{\pi}_i} \right); \\ b_1 &= \alpha_{10} + \gamma_1 \left(\sum_{i=1}^n \frac{X_i (1-Y_i)}{\hat{\pi}_i} \right); & b_0 &= \alpha_{01} + \gamma_0 \left(\sum_{i=1}^n \frac{(1-X_i) (1-Y_i)}{1-\hat{\pi}_i} \right); \\ \gamma_1 &= \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n X_i / \hat{\pi}_i}; & \gamma_0 &= \frac{\sum_{i=1}^n (1-X_i)}{\sum_{i=1}^n (1-X_i) / (1-\hat{\pi}_i)}. \end{aligned}$$

With $\hat{\pi}_i = \hat{\pi}(\mathbf{C}_i)$. The parameter a_1 is the prior number of subjects with events (α_{11}) plus the number of subjects with events in group 1 of the pseudo-population, and b_1 is the prior

number of subjects having no events (a_{10}) plus the number of subjects with no events in the group 1 pseudo-population. Both are renormalized by γ_1 so that their sum is equal to the size of the original (unweighted) group 1 population. Similar interpretations hold for a_0 , b_0 , and γ_0 in group 0. With the specification $\{a_{00}, a_{01}, a_{10}, a_{11}\} = \{1, 1, 1, 1\}$ the priors add very little information.

We can draw J times from the outcome model for each of K draws from the propensity score model. The j th draw from the estimated posterior distribution of the causal effect, Δ , using the k th draw for $\hat{\pi}^k$ can be calculated as:

$$\hat{\Delta}^{jk} | \mathbf{D}, \hat{\pi}^k(\mathbf{C}) = \hat{p}_1^{jk} - \hat{p}_0^{jk}. \quad (5)$$

Because (4) is conditional on a single value of $\hat{\pi}^k(\mathbf{C})$, we can evaluate $\Delta^{jk} | \mathbf{D}, \hat{\pi}^k(\mathbf{C})$ over the distribution of $\pi(\mathbf{C})$. Kaplan and Chen (2012) proposed combining the estimates using the law of total variance to get:

$$E(\hat{\Delta} | \mathbf{D}) = E_{\pi}\{E(\hat{\Delta} | \mathbf{D}, \hat{\pi})\} \text{ and } \text{Var}(\hat{\Delta} | \mathbf{D}) = E_{\pi}\{\text{Var}(\hat{\Delta} | \mathbf{D}, \hat{\pi})\} + \text{Var}_{\pi}\{E(\hat{\Delta} | \mathbf{D}, \hat{\pi})\}, \quad (6)$$

which resembles multiple imputation. Formula (6) demonstrates how the variance from each model contributes to the final uncertainty: the average variance from the outcome models enters as the first term in (6) and variance across the propensity scores enters as the second term. With our proposed algorithm we get draws from the outcome model for each propensity score draw, not just a variance estimate. Thus we recommend generating the entire posterior density $\Delta | \mathbf{D}$ by simply concatenating across propensity score draws. The variance can then be computed directly from these draws and uncertainty measures like 95% confidence intervals can be calculated from their empirical quantiles.

Saarela et al. (2015) critiqued the approach of Kaplan and Chen (2012) for using a non-likelihood outcome model when generating the posterior variance in equation (6). Because we specify a weighted binomial likelihood for the counts $\tilde{Y}(0)$ and $\tilde{Y}(1)$, our approach is grounded in the theoretical work of Saarela et al. (2015), who used propensity score weights along with a Dirichlet distribution for Bayesian bootstrapping to obtain posterior draws. However in contrast to the method of Saarela et al. (2015), our approach takes advantage of conjugacy and has clear connections to more conventional IPW.

In our supplementary materials, we show the equivalence between the posterior mean of equation (5) and the IPW estimator in equation (2). We also provide further computational details and R code for sampling from $\Delta | \mathbf{D}$.

5 Simulation Studies

5.1 Handling of Propensity Score Distribution

Saarela et al. (2015) made the case that even in a Bayesian framework, propensity scores should be fixed to their maximum likelihood estimates or posterior mean estimates rather than integrating over the full distribution. However, doing so in high-dimensions ignores the crucial source of uncertainty from estimating the propensity scores, which could potentially cause under coverage and result in more type I errors. To test this, we first constructed some very simple simulation scenarios. We let $X \sim \text{Bernoulli}\{(\text{logit}^{-1}(.5C_1 + .5C_2))\}$ and $Y \sim \text{Bernoulli}\{\text{logit}^{-1}(X - .5C_1 - .5C_2)\}$ where C_1 and C_2 were drawn from independent standard normals with $n = 100$ subjects. We specified three Bayesian logistic models to estimate the propensity scores. The first estimates the correct model, with C_1 and C_2 as confounders. The second estimates an “over specified” model that includes C_1 , C_2 , and 8 additional covariates with no relation to treatment or outcome. Thus there is a relatively low signal-to-noise ratio in the second scenario. Finally, we estimate an “under specified” model that includes only C_1 , making C_2 an unmeasured confounder. The results after 4000 simulations appear in table 2.

In their original paper, Saarela et al. (2015) focused on analyzing propensity score estimators across different levels of confounding. They found that including variance from the propensity score estimate tended to cause slight over coverage in all scenarios, while IPW and fixed propensity score Bayesian estimators had either perfect coverage or under covered the true parameter. We too find that integrating over the propensity score distribution leads to slight over coverage, making it slightly conservative. Fixing the propensity score parameter estimates to the posterior means leads to good coverage when there are few variables, but underestimates the variance in the over specified case and is slightly more biased. Because we are primarily concerned with over specified models in high-dimensions, in which there is likely to be a low signal-to-noise ratio, the robust coverage of the integrated method in this setting is key. Thus these initial simulations supported integrating over the propensity score distribution for high-dimensional analysis.

5.2 More Realistic Simulations

We compare three Bayesian methods to estimate the propensity score distribution: a logistic regression with weakly informative priors defined as Student- $t_3(0, 2.5)$ (Gelman et al., 2014), a logistic regression regularized by horseshoe priors, and BART with $m = 200$ trees grown using default priors on the tree depth and end nodes (details in appendix). We altered the horseshoe method slightly by replacing the half-Cauchy priors with half- $t_3(0, 1)$ priors, which allow much easier fitting using Monte Carlo sampling without substantive difference in estimates (Stan Development Team, 2016). A naive estimate is computed as the unadjusted difference in means between treatment groups. As a baseline competitor, we examine IPW as in equation (2) with empirical sandwich standard errors. We also compare to targeted maximum likelihood estimation (TMLE), a frequentist approach to causal inference in high-dimensional settings (van der Laan and Rose, 2011). TMLE used the SuperLearner ensemble method with unregularized and lasso penalized logistic regression as cross-validated candidate models to fit outcome and propensity score regressions. These

regressions are combined to formulate a doubly-robust estimate of the ATE (van der Laan and Rose, 2011).

We ran 500 simulations with $n = 1000$ and $p = 100$ constructed to partly resemble our coronary stent data in terms of coefficient values and confounder distributions, while remaining computationally tractable. To this end, 100 binary confounders were extracted from the data, fit to the treatment variable using logistic regression, and rounded to the nearest 10th. These estimated coefficients were then used as the true treatment coefficients (β^X) in simulated datasets. The values of β^X ranged from -1.1 to 1.1 on the log-odds scale, with a lower quartile of -0.2 and upper quartile of 0.2 , making most of the coefficients quite small, though only 18 were exactly equal to 0. The confounders C were drawn from independent Bernoulli distributions with probabilities equal to their empirical prevalence in our data. As a result, many of the covariates were quite sparse. The outcome coefficients β^Y were then selected to create confounding while enforcing $\sum \beta^Y = 0$. The data were simulated following: $P(X_i | C_i) \sim \text{Bernoulli}\{\text{logit}^{-1}(\beta_0^X + \sum_{j=1}^p \beta_j^X C_{ij})\}$, yielding a marginal treatment probability of $P(X_i = 1) \approx .7$, and $P(Y_i | X_i, C_i) \sim \text{Bernoulli}\{\text{logit}^{-1}(\beta_0^Y + \beta^{\text{Tr}} X_i + \sum_{j=1}^p \beta_j^Y C_{ij})\}$. We set $\beta^{\text{Tr}} = -2$, yielding a true marginal ATE of ≈ -0.15 . An outcome intercept of $\beta_0^Y = -2$ gave a relatively low event rate with $P(Y_i = 1) \approx 0.1$. We assessed bias, mean squared error (MSE), confidence interval width, and realized 95% posterior interval coverage of each method.

Weakly informative $t_3(0, 2.5)$ priors had almost perfect coverage, but the highest MSE of any method (Table 3). The lack of regularization likely made the treatment regressions unstable, and IPW suffered from similarly poor MSE results. The horseshoe and BART fared better, with considerably improved MSE, tight confidence intervals, and reasonably good coverage. BART even improved slightly over the horseshoe in terms of bias and coverage. It's competitive performance is remarkable because, unlike the logistic regression-based approaches which correctly specified the model as linear, BART is completely non-parametric.

Fixing the propensity scores to their posterior means considerably reduced variance estimates, as evidenced by tighter confidence intervals for all methods, but led to poor coverage as a result. Although the MSE of the $t_3(0, 2.5)$ priors improved, the MSE and bias were actually slightly higher for the other Bayesian methods. As expected, the naive difference in means performed very poorly. The IPW suffered from poor MSE and slightly under-covered the true treatment effect. While the TMLE had among the best performance with the lowest MSE, low bias, and tight confidence intervals, it under-covered the true treatment effect considerably, indicating the inadequacy of its standard error estimate. Bootstrapping TMLE would likely yield better coverage by incorporating uncertainty from estimating the scores, but was computationally infeasible in this simulation setting.

Our simulations show the effectiveness of Bayesian propensity scores and support including uncertainty from the propensity score distribution when fitting large treatment models. We also demonstrated the utility of Bayesian methods for regularization and non-parametric regression for reducing the MSE of high-dimensional causal estimators.

6 Application: Drug Eluting vs Bare Metal Coronary Stenting in Massachusetts

We revisit a comparison of the causal effect of using DES to BMS on the risks of all-cause mortality and a composite endpoint including mortality, acute myocardial infarction, and target vessel revascularization (TVR) within 1 year after the initial procedure. Past studies using propensity score techniques have consistently found a benefit to drug-eluting stents on lowering rates of 1-year mortality. In contrast RCTs and an instrumental variables analysis report no such benefit (Mauri et al., 2008) (Venkitachalam et al., 2011) (Bonaa et al., 2016). However, these analyses have utilized well fewer than 100 variables for confounding adjustment. An analysis using a richer confounder set may align more closely with experimental results. Although we are primarily interested in TVR, we use a composite endpoint, which has been used in past stent trials to avoid bias due to competing risks,

We examined $n = 8718$ patients from the Mass-DAC clinical registry consisting of percutaneous coronary interventions (PCIs) performed in Massachusetts adults between October 1, 2010 and September 30, 2011 inclusive. We merged the Mass-DAC data with hospital billing data from the Massachusetts Center for Health Information and Analysis (CHIA). The Mass-DAC data contributed 131 potential confounders, a mix of measures indicating patient characteristics (e.g. age, sex), pre-existing conditions (e.g. diabetes, heart failure), or characteristics of the procedure (e.g. treated vessel, hospital). There were 12 continuous covariates, the remainder being binary or categorical covariates. The CHIA data included 15 free-response fields containing Present On Admission (POA) diagnoses entered as International Classification of Diseases, 9th Version (ICD-9) codes. POA codes describe patients at hospital arrival and are not impacted by treatment decisions. We considered all diagnoses for which 10 or more patients were coded as having the condition and converted them to dummy variables. This yielded 364 additional covariates for a total of 495 potential confounders. Henceforth we refer to the combined dataset as Mass-DAC. We imputed a small amount of missing data, creating one imputed dataset for simplicity. Due to the small amount of missingness substantive inferences would be unlikely to change if we had used multiple imputed datasets.

We estimate the ATE using Bayesian propensity score weighting with treatment models estimated from the full Mass-DAC data. Our causal parameter is defined as the average event rate in the BMS group subtracted from the average event rate in the DES group. We focus on analyzing two outcomes: 1-year mortality and a composite endpoint of 1-year mortality, acute myocardial infarction (AMI), and TVR. We derived 1-year AMI from records of in-hospital events as mandated in the Mass-DAC registry, or from any within-year hospital readmission with diagnosis code 410.x1 in the CHIA data. We defined 1-year TVR as a PCI on the same vessel or any coronary artery bypass graft (CABG) surgery within 1 year of stent implantation. As an additional check, we include 30-day mortality as a falsifiability endpoint. Since a true causal effect on mortality is unlikely to manifest after 30-days, we know a significant difference on this outcome is attributable to residual confounding.

For the first stage we estimated Bayesian models with three different specifications: Student-t priors, horseshoe priors, and BART with 2000 trees grown using default parameters. The logistic regressions were fit using Stan and convergence was assessed by inspecting trace plots and by ensuring all within-chain relative to between-chain statistics (R-hat) were less than 1.1 (Stan Development Team, 2016). BART was fit in R using the package BayesTree, written by the original authors (Chipman et al., 2010). We implemented an IPW estimator as in equation (2) where propensity scores were estimated with logistic regression and the variance estimate used the empirical sandwich method or 2000 bootstrap replicates (Austin and Stewart, 2015). As a final competitor, we implemented TMLE with SuperLearner using unregularized and lasso penalized logistic regression as candidate models, with variance estimated from the influence curve or from 1000 bootstrap replicates (van der Laan and Rose, 2011). We denote the sandwich variance IPW and influence curve variance TMLE with the preface *analytic*, to reflect the closed-form nature of their variance estimates. We expect the analytic methods to have tighter confidence intervals because they do not reflect variance from the propensity score model.

We assess positivity, balance, and stability of our estimated propensity scores. We also examine the distributions of weights and the (weighted) absolute standardized differences in covariates between the DES and BMS groups (Austin and Stewart, 2015). High weights tend to result from an overfit propensity score model and can also indicate a positivity violation, though low weights are no guarantee of good overlap (Petersen et al., 2010). The standardized difference in covariates provides a measure of balance, where an absolute standardized difference of less than 10% is considered negligible (Austin and Stewart, 2015). For the Bayesian methods, we graphically examined the distributions of balanced standardized differences in order to assess post-weighting covariate balance as well as the stability of the estimates. Optimally, the 95% posterior intervals for the standardized difference would be within $(-10, 10)$ for all confounders, though in analogy to standard propensity score analyses balance is acceptable if all posterior means are within $(-10, 10)$ (Kaplan and Chen, 2012). Finally, we assessed the distribution of propensity score weights across subjects for both our Bayesian and non-Bayesian methods, defined as $X_j/\pi_j + (1 - X_j)/(1 - \pi_j)$. For the Bayesian methods the analyzed weights correspond to each subjects posterior mean weight, while for the frequentist methods there is only 1 weight (from the maximum likelihood propensity score) returned for each person.

6.1 Results

All Bayesian methods converged satisfactorily according to trace plots and R-hat statistics. All TMLE bootstrap iterations converged, though about 25% of the bootstrapped IPW estimates did not as a result of the data sparsity and lack of regularization.

Figure 1 shows the distribution of the percent standardized differences for each of the 495 confounders for the Bayesian propensity score methods, sorted by the mean under weakly informative $t_3(0, 2.5)$ priors. The $t_3(0, 2.5)$ priors for the logistic treatment regression coefficients show a significant portion of the standardized difference distributions for some confounders outside of the -10% to 10% range. In contrast, the standardized difference distributions under horseshoe priors for the logistic treatment regression coefficients and

those using a BART treatment model were almost all entirely within an acceptable range. Due to their heavier regularization, the posteriors under horseshoe priors and BART are more stable. The fact that some parts of the posterior under $t_3(0, 2.5)$ priors do not sufficiently balance the confounders is concerning. The $t_3(0, 2.5)$ priors had the highest average and maximum weights of all the approaches, while horseshoe priors and BART had the lowest. None of the methods' weights indicated a positivity violation.

Table 4 displays the point estimates and 95% posterior intervals for each method and outcome. In terms of the estimated causal effect, every method sharply decreased the magnitude of the naive 1-year (30-day) mortality difference from 6.9 (3.4) while also decreasing the magnitude of the composite endpoint difference from 9.4. Even after adjustment, most methods found a significant benefit to mortality for DES. This is particularly concerning for 30-day mortality as any differences observed that soon after implantation are almost certainly a result of residual confounding. Only bootstrapped IPW and Bayesian propensity scores with $t_3(0, 2.5)$ priors found no significant DES benefit to 30-day mortality, though the other methods found only a small benefit. Only the Bayesian logistic treatment model with $t_3(0, 2.5)$ priors found no benefit to 1-year mortality. These findings may reflect residual confounding, as past RCTs and instrumental variables analyses have repeatedly confirmed no mortality benefit (Venkitachalam et al., 2011) (Bonaa et al., 2016). However, our methods did adjust away considerably more confounding than Mauri et al. (2008), which also used the Mass-DAC data. As expected, all methods found a significant DES benefit to the composite endpoint, as shown in clinical trials.

BART provided covariate balance and estimates of the risk difference comparable to the horseshoe priors. In our stent example, accommodating non-linearities and interactions did not make a large difference on the causal estimate compared to the logistic models.

7 Discussion

The methods we discussed combine the advantages of a Bayesian framework and frequentist propensity scores for causal inference with many variables, where including as many potential confounders as possible can help meet the ignorability assumption. In particular, Bayesian logistic regression using Student- t priors or horseshoe priors and BART provide useful parametric and non-parametric options to estimating propensity score distributions from high-dimensional data sets. Through simulations we showed that integrating over the propensity score model can slightly reduce bias and mean squared error but significantly improve coverage in high-dimensional settings. Because a significant portion of the variance in a propensity score analysis stems from estimating the scores, incorporating uncertainty from the propensity score regression was important to achieving good coverage in our Bayesian approaches. We also found that IPW coupled with bootstrapping gave more reasonable inferential results in our application than the empirical sandwich method, which does not account for uncertainty in the propensity score model.

We proposed horseshoe priors and BART as techniques for variance reduction and flexible model fitting. While these approaches dominated more conventional Student- t priors in our simulations and looked better diagnostically in our application, they failed to meet our

falsifiability endpoint. For causal inference, we recommend avoiding regularization of well-established confounders, incorporating substantive knowledge for variable selection and prior specification when available, and using sparsity-inducing priors when there is noisy data. A Bayesian framework provides a comprehensive and intuitive way to build models that include substantive knowledge and regularization, or adaptive specifications when little is known about the data generating process.

In our application we assessed the benefits of DES over BMS on preventing mortality and a composite endpoint. We found that by including a rich set of almost 500 confounders for adjustment and using Bayesian propensity scores we were able to adjust for considerably more confounding than an earlier propensity score analysis with the same data (Mauri et al., 2008). When used with weakly informative priors our method aligned with estimates observed in RCTs (Bonna et al., 2016).

Given that a few approaches exist for causal inference in high-dimensional data settings we have several recommendations. First, although the propensity score is essentially a nuisance parameter, accounting for the variance in a high-dimensional treatment regression is important to achieving good coverage. Bootstrapping both stages of a propensity score analysis accomplishes this goal in a frequentist setting. Second, an outcome free design in which the treatment model is separated from the outcome model is often desirable and sometimes mandated. Third, making fewer parametric assumptions in the outcome model, for example by using a weighting procedure instead of regression adjustment, is advantageous. Our approach meets these criteria. The methods we discussed did not use any information from the outcome in setting up the design (treatment) model, allowing the user to focus on designing a good propensity score model without revealing the causal effect until balance has been achieved (Rubin, 2008). Such an approach is not without drawbacks, as including outcome information can aid variable selection or enable doubly-robust estimation (Schneeweiss et al., 2009)(Wang et al., 2012)(Robins et al., 1994)(Cefalu et al., 2016). Nevertheless, propensity scores remain a powerful and popular tool for causal inference and we believe that our proposed methods provide a useful framework for comparative effectiveness research with high-dimensional data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Mr. Spertus' and Dr. Normand's efforts were partially supported by a grant from the National Institute of General Medical Sciences R01-GM111339. Dr. Normand's effort was also partially supported by a grant from the Center for Devices and Radiological Health U01-FD004493. The authors are indebted to Ann Lovett, R.N., M.A. and Matt Cioffi, M.S., both from Harvard Medical School for preliminary dataset construction and review, and to the Massachusetts Department of Public Health and CHIA for the use of their data.

References

Austin PC and Stuart EA (2015). Moving toward best practice when using inverse probability of treatment weighting (IPW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine* 34, 3661–3679. [PubMed: 26238958]

- Austin PC (2010). The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine* 29, 2137–2148. [PubMed: 20108233]
- Bonaa KH, Mannsverk J, Wiseth R, Aaberge L, Myreng Y, Nygard O, et al. (2016). Drug-eluting or bare-metal stents for coronary artery disease. *The New England Journal of Medicine* 375, 1242–1252. [PubMed: 27572953]
- Cefalu M, Dominici F, Arvold N, and Parmigiani G (2016). Model Averaged Double Robust Estimation. *Biometrics* 73, 410–421. [PubMed: 27893927]
- Carvahlo CM, Polson NG, and Scott JG (2010). The horseshoe estimator for sparse signals. *Biometrika* 97, 465–480.
- Chipman HA, George EI, and McCulloch RE (2010). BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics* 4, 266–298.
- Farrell MH (2015). Robust inference for average treatment effects with possibly more covariates than observations. *Journal of Econometrics* 189, 1–23.
- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, and Rubin DB (2014). *Bayesian Data Analysis: Third Edition* CRC Press.
- Ghosh D (2011). Propensity score modeling in observational studies using dimension reduction methods. *Statistics and Probability Letters*, 81, 813–820. [PubMed: 21617766]
- Ghosh D, Zhu Y, and Coffman DL (2015). Penalized regression procedures for variable selection in the potential outcomes framework. *Statistics in Medicine*, 34, 1645–1658. [PubMed: 25628185]
- Graham DJ, McCoy EJ, and Stephens DA (2016). Approximate Bayesian Inference for doubly robust estimation. *Bayesian Analysis*, 11, 47–69.
- Hahn PR, Carvalho CM, He J, and Puelz D (2016). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*.
- Hahn PR, Murray JS, Carvalho CM (2017). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. <https://arxiv.org/abs/1706.09523>.
- Hill JL (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics* 20, 217–240.
- Kaplan D, and Chen J (2012). A two-step Bayesian approach for propensity score analysis: simulations and case study. *Psychometrika* 77, 581–609.
- Kass RE and Steffey D (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association* 84, 717–726.
- Lunceford JK, and Davidian M (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 23, 2937–2960. [PubMed: 15351954]
- Mauri L, Silbaugh TS, Wolf RE, Zelevinsky K, Lovett A, Zhou Z, Resnic FS, and Normand SLT (2008). Long-term clinical outcomes after drug-eluting and bare-metal stenting in Massachusetts. *Circulation* 118, 1817–1827. [PubMed: 18852368]
- McCaffrey DF, Ridgeway G, and Morral AR (2004). Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods* 9, 403–425. [PubMed: 15598095]
- Petersen ML, Porter KE, Gruber S, Wang Y, and van der Laan MJ (2010). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research* 21, 31–54. [PubMed: 21030422]
- Robins JM, Rotnitzky A, and Zhao LP (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89, 846–866.
- Rosenbaum PR and Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.
- Rubin DB (1985). The use of propensity scores in applied Bayesian inference. *Bayesian Statistics* 2, 463–472.
- Rubin DB (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics* 2, 808–840.

- Saarela O, Belzile LR, and Stephens DA (2016). A Bayesian view of doubly robust causal inference. *Biometrika* 103, 667–681.
- Saarela O, Stephens DA, Moodie EEM, and Klein MB (2015). On Bayesian estimation of marginal structural models. *Biometrics* 71, 279–301. [PubMed: 25677103]
- Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, and Brookhart AM (2009). High-dimensional propensity in studies of treatment effects using health care claims data. *Epidemiology* 103, 667–681.
- Stan Development Team. (2016). Stan modeling language users guide and reference manual, version 2.14.0. <http://mc-stan.org>
- U.S. Food and Drug Administration. (2013). Best practices for conducting and reporting pharmacoepidemiologic safety studies using electronic health records. <http://www.fda.gov/downloads/drugs/guidances/ucm243537.pdf>
- van der Laan MJ, and Rose S (2011). Targeted Learning: Causal Inference for Observational and Experimental Data Springer.
- Venkitachalam L, Lei Y, Magnuson E, Chan PS, Stolker JM, Kennedy KF, Kleiman NS, and Cohen DJ (2011). Survival benefit with drug-eluting stents in observational studies: fact or artifact? *Circulation* 124, 587–594. [PubMed: 21988921]
- Wang C, Parmigiani G, and Dominici F (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* 68, 661–686. [PubMed: 22364439]
- Zigler CM (2016). The Central Role of Bayes Theorem for Joint Estimation of Causal Effects and Propensity Scores. *The American Statistician* 70, 47–54. [PubMed: 27482121]
- Zigler CM and Dominici F (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association* 109, 95–107. [PubMed: 24696528]

Figure 1:

Mean (black circle) and 95% posterior interval (grey bar) of standardized difference for each confounder after balancing on each draw from the propensity score model. Each panel represents a different propensity score model specification. The dotted lines at -10 and 10 indicate range usually taken to be acceptable. Points are sorted by the mean standardized difference from the weakly informative $t_3(0, 2.5)$ specification.

Table 1:

Prevalence or mean (SD) of selected confounders and outcomes in Mass-DAC data. TVR = target vessel revascularization; STEMI = ST-elevated myocardial infarction; LVSD = left ventricular systolic dysfunction.

Characteristic	Stent Type		
	Overall	Bare Metal	Drug Eluting
Outcomes			
1-Year Mortality, %	5.7	10.2	3.3
1-Year TVR, %	7.4	9.0	6.5
Confounders			
Mean Age (SD), yrs	64.7 (12.5)	66.4 (11.7)	63.7 (13.5)
STEMI, %	24.4	35.7	18.2
Cardiomyopathy or LVSD, %	9.4	11.1	8.4
Emergent Status, %	26.7	38.3	20.3
Cardiogenic Shock, %	1.8	3.8	0.8

Table 2:

Properties of Bayesian propensity score weighting over 4000 simulations where propensity scores are either integrated over their posterior distribution or fixed to their posterior means. Sample size within each simulation is $n = 100$. Coverage of 95% posterior intervals is reported.

		Integrated PS	Mean PS
Correct	Bias	.0008	-.0069
	Variance	.0102	.0086
	Coverage	96.5%	94.7%
		Integrated PS	Mean PS
Overspecified	Bias	.0006	-.0270
	Variance	.0157	.0085
	Coverage	96.6%	91.1%
		Integrated PS	Mean PS
Underspecified	Bias	-.0519	-.0566
	Variance	.0094	.0087
	Coverage	91.9%	90.4%

Table 3:

Bias (estimate - true effect), mean squared error times 1000 ($MSE \times 10^3$), average confidence interval (CI) width, and coverage of 95% posterior intervals for various estimators over 500 simulations with sample size $n = 1000$ and $p = 100$ confounders specified above. Bayesian propensity scores were implemented with different treatment of the propensity score distribution (integrated or mean), and different modeling choices (horseshoe priors, $t_3(0, 2.5)$, or Bayesian additive regression trees (BART)). A naive difference in means, classic inverse probability weighting with robust standard errors (IPW), and targeted maximum likelihood (TMLE) were also evaluated.

Treatment Model	Bias	MSE $\times 10^3$	CI Width	Coverage
<i>Integrated Propensity Score</i>				
Logistic, $t_3(0, 2.5)$ Priors	-.011	2.82	.220	95.2%
Logistic, Horseshoe Priors	.016	0.93	.110	93.0%
BART, Default Regularizing Priors	.011	0.84	.123	96.8%
<i>Mean Propensity Score</i>				
Logistic, $t_3(0, 2.5)$ Priors	-.001	1.59	.095	79.2%
Logistic, Horseshoe Priors	.018	0.96	.092	86.0%
BART, Default Regularizing Priors	.015	0.86	.093	87.2%
<i>Other Methods</i>				
IPW	-.001	1.67	.151	92.8%
TMLE	.006	0.78	.075	81.4%
Naive Estimate	.030	1.46	.092	73.0%

Table 4:

Mean (max) percent absolute standardized difference (Std Diff) across covariates, mean (max) propensity score weights, estimates of causal effects (95% confidence interval) as percent difference in outcome rate between DES and BMS, $\hat{\Delta} = \hat{p}_1 - \hat{p}_0$, after adjustment. A negative estimate indicates DES yields a lower rate than BMS. IPW = inverse probability weighting; TMLE = targeted maximum likelihood.

Approach	Weight	Mortality		
		Mortality 30-day	1-Year	1-Year Composite
Naive	NA	-3.4 (-4.1, -2.6)	-6.9 (-8.0, -5.7)	-9.4 (-11.1, -7.6)
Analytic TMLE	NA	-1.1 (-1.7, -0.6)	-3.0 (-4.0, -2.0)	-7.3 (-9.0, -5.5)
Bootstrapped TMLE	NA	-1.1 (-2.7, -0.5)	-3.0 (-4.4, -1.8)	-7.3 (-10.0, -5.2)
Analytic IPW	2.0 (54.9)	-1.0 (-2.0, -0.1)	-2.9 (-4.3, -1.5)	-6.9 (-9.3, -4.4)
Bootstrapped IPW	2.0 (54.9)	-1.0 (-2.0, 1.8)	-2.9 (-4.9, -0.3)	-6.9 (-10.1, -3.1)
Logistic, $\mathcal{L}(0,2.5)$ Priors	2.3 (100.6)	-0.3 (-1.7, 2.5)	-2.1 (-4.2, 0.9)	-6.1 (-9.3, -2.4)
Logistic, Horseshoe Priors	2.0 (29.6)	-1.2 (-2.0, -0.5)	-3.2 (-4.5, -2.0)	-7.6 (-9.6, -5.5)
BART, Default Regularizing Priors	2.0 (28.2)	-1.3 (-2.1, -0.6)	-3.3 (-4.5, -2.1)	-7.4 (-9.3, -5.5)