Biometrics 69, 263–273
March 2013

DOI: 10.1111/j.1541-0420.2012.01830.x

Model Feedback in Bayesian Propensity Score Estimation

Corwin M. Zigler,[1,*] Krista Watts,[1] Robert W. Yeh,[2] Yun Wang,[1] Brent A. Coull,[1] and Francesca Dominici[1]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.
[2]Cardiology Division, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston Massachusetts 02114, U.S.A.
[*]*email*: czigler@hsph.harvard.edu

Summary. Methods based on the propensity score comprise one set of valuable tools for comparative effectiveness research and for estimating causal effects more generally. These methods typically consist of two distinct stages: (1) a propensity score stage where a model is fit to predict the propensity to receive treatment (the propensity score), and (2) an outcome stage where responses are compared in treated and untreated units having similar values of the estimated propensity score. Traditional techniques conduct estimation in these two stages separately; estimates from the first stage are treated as fixed and known for use in the second stage. Bayesian methods have natural appeal in these settings because separate likelihoods for the two stages can be combined into a single joint likelihood, with estimation of the two stages carried out simultaneously. One key feature of joint estimation in this context is "feedback" between the outcome stage and the propensity score stage, meaning that quantities in a model for the outcome contribute information to posterior distributions of quantities in the model for the propensity score. We provide a rigorous assessment of Bayesian propensity score estimation to show that model feedback can produce poor estimates of causal effects absent strategies that augment propensity score adjustment with adjustment for individual covariates. We illustrate this phenomenon with a simulation study and with a comparative effectiveness investigation of carotid artery stenting versus carotid endarterectomy among 123,286 Medicare beneficiaries hospitlized for stroke in 2006 and 2007.

Key words: Bayesian estimation; Causal inference; Comparative effectiveness; Model feedback; Propensity score.

1. Introduction

One valuable class of methods for comparing the effectiveness of clinical treatments as they are applied in routine practice relies on the notion of the propensity score (PS; Rosenbaum and Rubin, 1983) to estimate causal effects that are not confounded by observed characteristics. Estimating causal effects with PS methods is achieved in two stages: (1) a "PS stage" where a model is fit to predict the receipt of treatment from available covariates, with the predicted values from this model representing the estimated PS, and (2) an "outcome stage" whereby outcomes of treated and untreated units are compared among units with similar values of the PS. Typically, the two-stage nature of the problem is accommodated by separate and sequential estimation; a model is fit in the PS stage, then the estimated PS from this model are treated as fixed and known to conduct adjusted comparisons in the outcome stage.

Only recently has Bayesian estimation been proposed as a means to jointly estimate quantities in the PS and outcome stages (McCandless, Gustafson, and Austin, 2009). One major motivation for Bayesian PS estimation is that jointly estimating quantities in the two stages propagates uncertainty in estimation of the PS into estimation of the treatment effect, whereas one well-known limitation of traditional sequential methods is that they potentially misstate the uncertainty in causal estimates by treating the estimated PS as a known quantity in the outcome stage (Gelman and Hill, 2007). The key idea with joint Bayesian PS estimation is that the PS is acknowledged as an unknown quantity, uncertainty about which is integrated out of posterior distributions of quantities in the outcome stage. Aside from providing a more comprehensive account of uncertainty, clear potential lies in incorporating PS methods into the broader literature on Bayesian methodology.

One salient feature of unifying distinct modeling stages with Bayesian estimation is that doing so allows "feedback" between the stages. In the PS context, this means that posterior samples of parameters in the PS stage are informed in part by information from the outcome stage, rendering the problem of Bayesian PS estimation substantially more nuanced than a simple Bayesian analog to well-established procedures. In fact, the notion of estimation and use of the PS in a joint likelihood has generated some controversy. One view is that the PS is meant to approximate the design stage of a randomized study, and that this should be done without any access to the outcome to ensure objective design decisions that are completely separate from analysis decisions (Rubin, 2007, 2008). Nonetheless, methods that incorporate outcome

information have been advocated (McCandless et al., 2009; Schneeweiss et al., 2009). In principle, incorporating feedback in joint Bayesian PS estimation entails estimates of the PS themselves that make more complete use of the data, which could improve estimation of causal effects. However, a rigorous investigation of exactly how feedback can impact estimation of causal effects is lacking.

In what follows we illustrate that, in general, model feedback in joint Bayesian PS estimation can result in biased estimates of the treatment effect. Unlike traditional sequential procedures that estimate the PS using only information on how covariates relate to the treatment, we show that joint Bayesian PS estimation with feedback uses information from the outcome model to construct the PS, and that this type of feedback can distort the nature of the PS and impair its ability to adjust for confounding. We also show that the nature of feedback is changed when using outcome models that augment PS adjustment with adjustment for individual covariates, and that this strategy can recover causal effects.

Using nationwide data on 123,286 Medicare beneficiaries, we illustrate joint Bayesian PS estimation in a comparative effectiveness investigation regarding the recent increase in the use of carotid artery stenting (CAS) for treatment of carotid artery disease (a primary cause of stroke), as compared to the more established carotid endarterectomy (CEA) procedure. Because these therapies are not randomly applied in clinical practice, we make use of numerous clinical characteristics to adjust for confounding in pursuit of a causal treatment effect estimate. We compare the results of the joint Bayesian analysis with a traditional sequential approach.

Section 2 of this paper briefly reviews PS estimation and outlines a traditional sequential estimation approach. Section 3 provides the details of joint Bayesian PS estimation and offers a transparent exploration of the role of feedback. Section 4 provides a simulation study to illustrate the role of feedback in comparison with conventional sequential methods. Section 5 uses Medicare data to compare the effectiveness of CEA versus CAS for preventing mortality within the first year of hospitalization. We conclude with a discussion.

2. Propensity Score Estimation

For a binary treatment, $X = 0, 1$, an outcome, Y, and a vector of p covariates $(C_1, C_2,...,C_p)$, Rosenbaum and Rubin (1983) defined the PS as the conditional probability of assignment to treatment $X = 1$, given the covariates. Causal inference with the PS relies on two important features. First, treatment assignment must be assumed strongly ignorable, that is, there must be no unmeasured confounders. Second, by virtue of the fact that the PS reflects the treatment assignment mechanism, the PS enjoys the property of a *balancing score*, resulting in conditional independence between the treatment and the individual covariates, conditional on the score: $X \perp\!\!\!\perp C_1, \ldots, C_p | \text{PS}$. This balancing score property combined with the assumption of strongly ignorable treatment assignment allows average comparisons between treated and untreated outcomes at a given value of the PS to serve as an unbiased estimate of the average treatment effect at that value of the PS.

2.1 *Models for the PS and Outcome Stages*

PS methods consist of two distinct stages: the PS stage that estimates a model for the treatment assignment mechanism, and the outcome stage that uses the PS to compare outcome values between treated and untreated units with similar covariate characteristics. The PS stage consists of a model for the probability that $X = 1$ (given covariates): $g_x\{E(X|C)\} = C\gamma$, where $g_x(\cdot)$ is a link function, and C is the collection of pretreatment covariates plus an intercept, $C = (1, C_1, C_2, ...,C_p)$. Thus, the PS stage can be represented with the following likelihood:

$$L(\mathbf{X}|\gamma, \mathbf{C}) = \prod_{i=1}^{n} \left\{g_x^{-1}(C_i\gamma)\right\}^{X_i} \left\{1 - g_x^{-1}(C_i\gamma)\right\}^{1-X_i}, \quad (1)$$

where, here and throughout, boldface is used for vectors and matrices representing the values for the entire sample, and $i = 1, \ldots, n$ indexes observational units. With this formulation, the values of γ and C_i determine the PS $\equiv P(X_i = 1|C_i)$ for the i^{th} unit.

Consider a binary outcome, $Y = 0, 1$, but note that results in the following analogously hold for other outcomes (e.g., continuous or survival). We confine our attention to settings that posit a model for the outcome, conditional on the PS: $g_y\{E(Y|X, C)\} = \xi_0 + \beta X + \xi h(\gamma, C) + C^+\delta$, where $g_y(\cdot)$ is another link function, the deterministic function $h(\gamma, C)$ specifies how the PS enters the outcome model, and the term $C^+\delta$ denotes possible residual adjustment for some subset $C^+ \in C$ in addition to the PS. For example, $h(\gamma, C) = C\gamma$ would specify linear adjustment for the link-transformed PS from (1), and $\delta = 0$ would indicate adjustment for the transformed PS only. Alternatively, $h(\gamma, C)$ could specify dummy variables for membership in subclasses defined by q quantiles of the PS, and $\delta \neq 0$ could augment PS adjustment with individual covariate adjustment within subclass. We express the outcome stage likelihood as:

$$\begin{aligned} L(\mathbf{Y}|\beta, \xi, \mathbf{X}, \mathbf{C}, \gamma, \delta) = \prod_{i=1}^{n} &\left\{g_y^{-1}(\xi_0 + \beta X_i + \xi h(\gamma, C_i) \right. \\ &\left. + \delta C_i^+)\right\}^{Y_i} \left\{1 - g_y^{-1}(\xi_0 + \beta X_i \right. \\ &\left. + \xi h(\gamma, C_i) + \delta C_i^+)\right\}^{1-Y_i}. \end{aligned} \quad (2)$$

The primary objective is to estimate the causal effect of $X = 1$ versus $X = 0$ on Y. Toward this end, the conditional parameter β may be of primary interest, but issues such as non-collapsibility imply that this conditional parameter may not describe the marginal causal effect in the population (Greenland, Robins, and Pearl 1999). Nonetheless, much of the subsequent equates estimation of causal effects to estimation of β for ease of illustration, as marginalizing over covariate distributions to obtain the marginal effect would also require adequate estimation of β as a precursor step.

2.2 *Traditional Sequential Estimation*

Traditional PS procedures conduct estimation in the PS and outcome stages completely separately. Estimates of γ are obtained from (1) to construct the estimated PS. Then, the estimated PS are treated as known and used in the outcome

model. That is, with estimated $\widehat{\gamma}$, estimation of the treatment effect follows from $L(\mathbf{Y}|\beta,\xi,\mathbf{X},\mathbf{C},\widehat{\gamma},\delta)$ specified in (2).

An important feature of this approach is that it makes no attempt to recover the entire covariate-outcome response surface. Rather than specify a model for the relationship between each covariate and the outcome, the outcome model conditions on a one-dimensional summary of multivariate covariate information (the PS), with the dimension reduction determined by fitting the PS model in (1). Of key importance is that this dimension reduction reflects the treatment assignment mechanism to ensure the balancing-score property. Other dimension-reductions of C that imply different values of γ may fail to reflect $p(X=1|C)$, and are not guaranteed to entail the balancing-score property at the heart of PS methods.

With sequential estimation, estimates of γ from (1) are obtained in a manner completely agnostic with regard to quantities in the outcome model such as β, ξ, and Y. As we elaborate in the following sections, the primary difference with joint Bayesian PS estimation is the presence of feedback, which means that specification of the outcome model affects estimates of γ.

3. Bayesian Estimation and Model Feedback

In this section, we formalize Bayesian PS estimation and illuminate in detail the role of model feedback. In contrast to the sequential procedure described in Section 2.2, Bayesian PS estimation combines the models in (1) and (2) into a single joint likelihood:

$$L(\mathbf{Y},\mathbf{X}|\mathbf{C},\gamma,\beta,\xi,\delta)=\prod_{i=1}^{n}\left\{g_x^{-1}(C_i\gamma)\right\}^{X_i}\left\{1-g_x^{-1}(C_i\gamma)\right\}^{1-X_i}\times \tag{3}$$

$$\left\{g_y^{-1}(\xi_0+\beta X_i+\xi h(\gamma,C_i)+\delta C_i^{+})\right\}^{Y_i}\left\{1-g_y^{-1}(\xi_0+\beta X_i +\xi h(\gamma,C_i)+\delta C_i^{+})\right\}^{1-Y_i}. \tag{4}$$

The likelihood in (3)-(4), coupled with prior distributions for γ,β,ξ, and δ serves as the basis for posterior inference. Recall that $h(\gamma,C)$ is a deterministic function of γ, which means that the PS themselves are treated as unknown quantities that are updated with every posterior update of γ. Model feedback in this case arises because γ appears in both terms of the likelihood, leading to posterior samples of γ that involve both the PS model and the outcome model.

Throughout, we use a Metropolis–Hastings MCMC algorithm to sample from posterior distributions. We conduct the MCMC using two sampling blocks: one updating γ from its conditional posterior distribution, which implies a corresponding update of the PS, and another block updating all parameters in the outcome model. Note from the likelihood in (3)-(4) that updating γ conditional on (β,ξ,δ) corresponds to an update of the PS and will involve both terms of the likelihood, but updating (β,ξ,δ) conditional on γ will only involve term (4) pertaining to the outcome model.

To illustrate the fundamental features of feedback implied by joint estimation of (3)-(4), the remainder of Section 3 considers the simplified setting where the outcome model entails linear adjustment for $g_x(\text{PS})$, that is, in (4) we assume that $h(\gamma,C)=C\gamma$ and that ξ is a scalar, ξ_1.

3.1 *Algebraic Illustration of Feedback*

Purely for illustration, take $g_x^{-1}(\cdot)$ and $g_y^{-1}(\cdot)$ as the Normal CDF, $\Phi(\cdot)$, representing Probit regression in the PS and outcome stages, and take all prior distributions $\propto 1$. Following Albert and Chib (1993), the Probit link allows Bayesian estimation with a data-augmentation procedure that iteratively samples Normally distributed latent continuous data with unit variance such that the latent $X^*(Y^*)$ are >0 when $X=1(Y=1)$, and <0 otherwise. Conditional on simulated $(\mathbf{X}^*,\mathbf{Y}^*)$,

$$p(\gamma,\beta,\xi,\delta|\mathbf{X}^*,\mathbf{Y}^*,\mathbf{X},\mathbf{Y},\mathbf{C})\propto\exp[-0.5\{(\mathbf{X}^*-\mathbf{C}\gamma)^{\mathbf{T}}(X^*-C\gamma)+\widetilde{\mathbf{Y}}^{\mathbf{T}}\widetilde{\mathbf{Y}}\}], \tag{5}$$

where $\widetilde{\mathbf{Y}}=(\mathbf{Y}^*-\xi_0\mathbf{1_n}-\beta\mathbf{X}-\xi_1\mathbf{C}\gamma-\mathbf{C}^{+}\delta)$, $\mathbf{C}$ is the $n\times(p+1)$ design matrix, and $\mathbf{1_n}$ is a $n-$dimensional vector with every entry equal to one. Thus, the conditional posterior distribution of γ can be written as:

$$p(\gamma|\mathbf{X}^*,\mathbf{Y}^*,\mathbf{X},\mathbf{Y},\mathbf{C},\beta,\xi,\delta)\propto\exp\left[\gamma^T\left\{\mathbf{C}^{\mathbf{T}}\mathbf{C}\left(1+\xi_1^2\right)\right\}\gamma-2\gamma^T\left\{\mathbf{C}^{\mathbf{T}}(\mathbf{X}^*+\xi_1(\mathbf{Y}^*-\xi_0\mathbf{1_n}-\beta\mathbf{X}-\mathbf{C}^{+}\delta))\right\}\right], \tag{6}$$

which corresponds to the kernel of a Normal distribution with covariance matrix $\{\mathbf{C}^{\mathbf{T}}\mathbf{C}(1+\xi_1^2)\}^{-1}$ and mean $\{\mathbf{C}^{\mathbf{T}}\mathbf{C}(1+\xi_1^2)\}^{-1}[\mathbf{C}^{\mathbf{T}}\{\mathbf{X}^*+\xi_1(\mathbf{Y}^*-\xi_0\mathbf{1_n}-\beta\mathbf{X}-\mathbf{C}^{+}\delta)\}]$. Immediately we see that when $\xi_1\neq 0$, quantities from the outcome model impact posterior estimates of γ and, by extension, the PS. This is the nature of model feedback.

3.2 *Implied Parameterization of the Covariate-Outcome Response Surface*

Until otherwise noted, assume an outcome model that only adjusts for the PS, that is, assume $\delta=0$ in (4). Considering the joint likelihood in (3)-(4) implies a parameterization of the covariate-outcome response surface. We re-express $\xi_0+\beta X+\xi h(\gamma,C)$ from term (4) as:

$$\begin{aligned}&\xi_0+\beta X+\xi_1(\gamma_0+\gamma_1C_1+\cdots+\gamma_pC_p)\\&\quad=(\xi_0+\xi_1\gamma_0)+\beta X+\xi_1\gamma_1C_1+\cdots+\xi_1\gamma_pC_p.\end{aligned} \tag{7}$$

The key feature of model feedback is that posterior estimates of γ are informed in part by this parameterization of the outcome model, which may imply information about γ that is not consistent with the treatment assignment mechanism. In particular, this will occur if the true covariate-outcome response surface cannot be expressed by rescaling the covariate-treatment surface (characterized by γ) by a single scalar, namely, ξ_1 in (7).

To further illustrate, consider a simple setting where the true underlying relationships between p covariates, treatment, and outcome is described as follows:

$$g_x\{P(X_i=1|C_i)\}=\gamma_0+\gamma_1C_{i1}+\cdots+\gamma_pC_{ip}\quad\text{and} \tag{8}$$

$$g_y\{P(Y_i=1|X_i,C_i)\}=\alpha_0+\beta X_i+\alpha_1C_{i1}+\cdots+\alpha_pC_{ip}. \tag{9}$$

With the above data-generating mechanism, the joint likelihood in (3)-(4) with $\delta=0$ correctly models (8), but entails linear adjustment for $g_x(PS)$, rather than a model for the complete covariate-outcome response surface in (9). Combining

the above data-generating mechanism with the paramaterization of expression (7) corresponds to $\alpha_0 = (\xi_0 + \xi_1\gamma_0)$ and $\alpha_1 = \xi_1\gamma_1, \alpha_2 = \xi_1\gamma_2, \ldots, \alpha_p = \xi_1\gamma_p$, meaning that the only way that the PS and outcome modeling stages can imply the same values of γ is if $\alpha_k = \xi_1\gamma_k$ for all k. If this relationship does not hold, then feedback from the outcome model will yield posterior estimates of γ that do not reflect the true treatment-assignment mechanism in (8), meaning that $h(\gamma, C)$ is not technically a function of the PS and may not be a balancing score. Thus, Bayesian estimation with (3)-(4) and $\delta = 0$ is not guaranteed to yield estimates of β that reflect the causal treatment effect. In contrast, the sequential strategy in Section 2.2 estimates γ without regard to the outcome model, thus ensuring that $h(\gamma, C)$ maintains the balancing-score property. We illustrate this phenomenon in the simulation study of Section 4.

3.3 *Augmenting PS Adjustment with Individual Covariates*

The above feature of joint Bayesian PS estimation is not a feature of model feedback in general, but rather a byproduct of the dimension reduction implied by using the PS as a univariate summary of covariate information. Consider instead a model with $\delta \neq 0$ that adjusts for covariates in addition to the PS. With $h(\gamma, C) = C\gamma$, C^+ can include at most $(p-1)$ covariates to prevent perfect multicolinearity. In this case, setting $C^+ = (C_2, \ldots, C_p)$, the right hand side of expression (7) denoting the implied parameterization of the covariate-outcome response surface becomes,

$$(\xi_0 + \xi_1\gamma_0) + \beta X + \xi_1\gamma_1 C_1 + (\xi_1\gamma_2 + \delta_1)C_2 + \cdots + (\xi_1\gamma_p + \delta_{p-1})C_p. \quad (10)$$

Thus, setting $\delta \neq 0$ allows the additional flexibility of modeling the covariate-outcome response surface without assuming a univariate rescaling of the covariate-treatment surface. Although setting $\delta \neq 0$ still implies feedback, the feedback does not imply the same restriction on the relationship between the covariate-treatment and covariate-outcome surfaces, which allows estimation of γ in accordance with the treatment assignment mechanism, thus maintaining the balancing-score property. The simulation study in Section 4 also illustrates this phenomenon, and the discussion draws connections with the notion of "double robustness."

4. Simulation Study of Feedback in joint Bayesian PS Estimation

In this section we present a simulation study to illustrate that the features described in the simplified setting of Section 3 persist in settings with more flexible specification of $h(\gamma, C)$. All simulated data sets contain $n = 1000$ observations and $p = 6$ covariates, simulated from the following data-generating scheme. First, $C_1, \ldots, C_6$ are simulated from a multivariate normal distribution with mean $(0, 0, 0, 0, 0, 0)$ and the identity covariance matrix. For all i, X_i is simulated from a Bernoulli distribution with:

$$P(X_i = 1|C_i) = \frac{\exp(\gamma_0 + \gamma_1 C_{i1} + \cdots + \gamma_6 C_{i6})}{1 + \exp(\gamma_0 + \gamma_1 C_{i1} + \cdots + \gamma_6 C_{i6})}. \quad (11)$$

All Y_i are similarly generated from Bernoulli distributions with:

$$P(Y_i = 1|X_i, C_i) = \frac{\exp(\alpha_0 + \beta X_i + \alpha_1 C_{i1} + \cdots + \alpha_6 C_{i6})}{1 + \exp(\alpha_0 + \beta X_i + \alpha_1 C_{i1} + \cdots + \alpha_6 C_{i6})}. \quad (12)$$

The values of γ specify the true treatment assignment mechanism, those of α specify the true covariate-outcome response surface, and β is the conditional treatment effect. For all simulations, we set $\beta = 0.0$.

We simulated 1000 data sets under each scenario described below, and analyze the simulated data with the joint Bayesian method described in Section 3. For comparison, we obtain maximum likelihood estimates of β using the traditional sequential procedure of Section 2.2 and from fitting model (12) directly, referring to the latter as the "Gold Standard" because we know that this is the true data-generating mechanism.

Throughout analysis of the simulated data, we specify both $g_x^{-1}(\cdot)$ and $g_y^{-1}(\cdot)$ as $\frac{\exp(\cdot)}{1+\exp(\cdot)}$, indicating logistic regression in both model stages. Unlike the simple illustrations provided in Section 3, we take a more flexible modeling approach that stratifies units on quintiles of the logit(PS) and estimates the same β across PS strata. Adjustment for PS subclass is augmented with additional covariate adjustment ($\delta \neq 0$) when noted. For the joint Bayesian PS analysis, every posterior update of γ implies an update of the PS, so the quintiles of logit(PS) are recalculated and the PS subclasses redefined at every MCMC iteration. We specify diffuse prior distributions for all parameters as Normal with mean 0 and variance 10^{10}. In addition to comparing estimates of β, we also compare methods on the basis of estimates of γ, which imply the estimated PS. For point estimation, we use posterior mean estimates for the joint Bayesian method, obtained from three MCMC chains, each run for 10,000 iterations, with the first 5000 discarded as burn in and every 10th sample saved for posterior inference. Note here that application of PS methods in practice should involve an investigation of whether covariates are balanced within PS subclass, which we forego in the simulation study. Balance checks are addressed in detail for the data analysis in Section 5.

4.1 *Scenario where the Covariate-Outcome Response Surface is a Simple Rescaling of the Covariate-Treatment Surface*

Scenario 1 generates data with parameters in (11) and (12) set to $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6) = (0.0, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3)$ and $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = (0.0, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$. This scenario represents a unique special case where a univariate rescaling of the covariate-treatment surface correctly characterizes the covariate-outcome response surface ($\alpha_k = \xi_1\gamma_k$ for $k = 1, \ldots, p$).

We analyze the data with $\delta = 0$. Figure 1 depicts boxplots of the resulting posterior estimates of γ and β, along with estimates from the sequential approach. We see that, on average, both methods produce point estimates of γ that are similar and agree with the true parameter values from (11). For $\gamma_1, \ldots, \gamma_6$, point estimates are less variable with the joint Bayesian method, which is to be expected because posterior distributions of these quantities involve additional information via feedback from the outcome model. Estimates of β

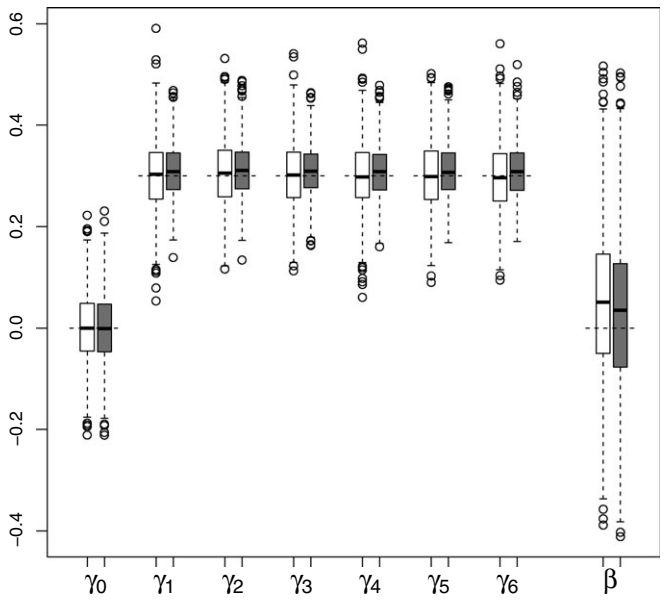


Figure 1. Scenario 1 with $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6) = (0.0, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3)$, and $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = (0.0, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$: boxplots of estimates of γ and β from the sequential frequentist and joint Bayesian analysis of 1000 replicated data sets. Shaded boxes are for the joint Bayesian analysis, unshaded are for the traditional sequential analysis. Horizontal dotted lines are at the true parameter values.

are similar between the two methods. This simulation illustrates the special case where the PS and outcome models imply the same values of γ, so posterior estimates of $h(\gamma, C)$ reflect the treatment assignment mechanism and the joint Bayesian method estimates the causal effect.

4.2 *Scenario where One Covariate Exhibits a Different Covariate-Treatment Relationship*

Appealing to the discussion in Section 3, we simulate Scenario 2 with 5 covariates having the same covariate-treatment and covariate-outcome relationships, with the sixth covariate exhibiting a different covariate-treatment relationship. This setting illustrates the effect that model feedback can have on joint Bayesian PS estimation when the covariate-outcome response surface cannot be expressed as a simple rescaling of the covariate-treatment surface. Toward this end, we simulate data as in Scenario 1, except we change γ_6 to -0.3 so that every γ_k cannot be rescaled by the same factor to obtain α_k.

We first analyze the data with $\delta = 0$. Figure 2(a) depicts boxplots of estimates of γ and β from both estimation methods. Unlike in Scenario 1, we see that, on average, the two estimation methods produce different estimates of $\gamma_1, \ldots, \gamma_6$. Whereas the traditional sequential approach estimates γ in accordance with the treatment assignment mechanism in (11), the joint Bayesian method estimates different values of γ, with the most pronounced difference is evident for γ_6. Thus, in the joint Bayesian method, the quantity $h(\gamma, C)$ does not reflect the treatment assignment mechanism, and is not guaranteed to serve as a balancing score. The result is posterior estimates of β with poor performance relative to estimates from the sequential procedure. This illustrates how feedback can distort the balancing-score property of the PS and yield estimates of β that do not reflect a causal effect.

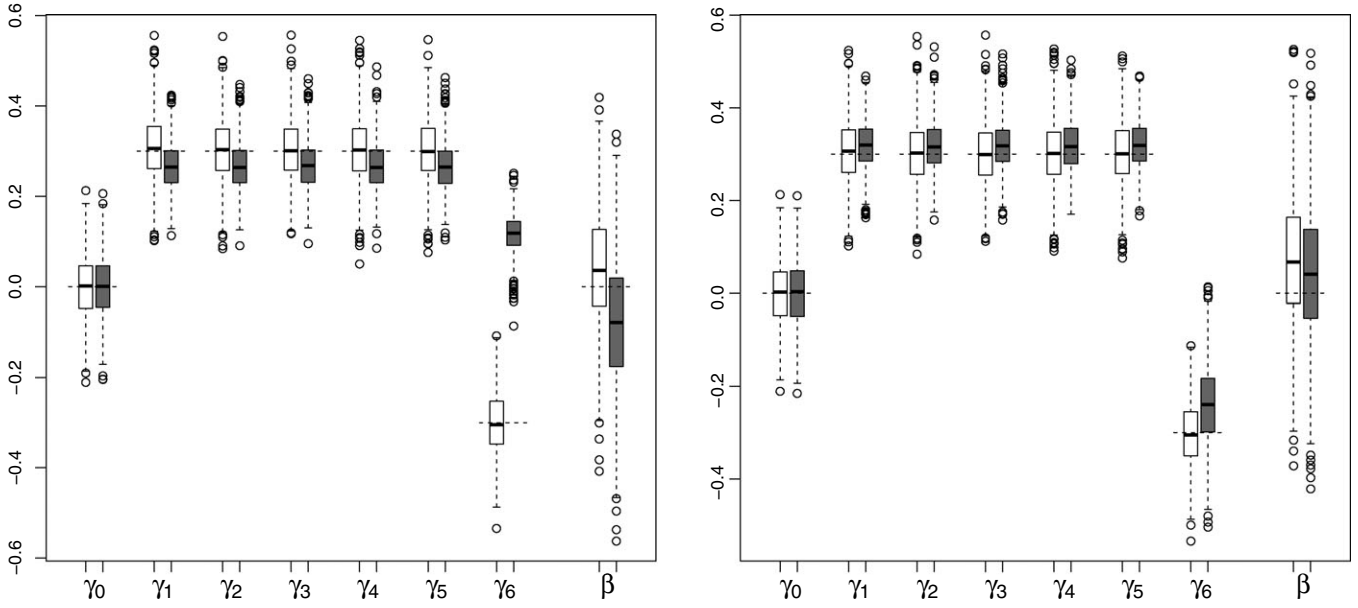
We argued in Section 3.3 that augmenting PS adjustment with individual covariates can prevent feedback from distorting estimates of γ. Because we know in this simulated example that one covariate exhibits a different relationship with the treatment, we reanalyze these simulated data sets with an outcome model that adjusts for C_6 within PS subclass. That is, we let $\delta \neq 0$ and $C^+ = C_6$ in (2), referring to this analysis as Scenario 2^+. Point estimates from this analysis are compared in Figure 2(b). Compared to the analysis that adjusts only for the PS, the model that augments PS estimation with additional adjustment of C_6 produces estimates of $\gamma_1, \ldots, \gamma_6$ that are much more similar between the two estimation methods, implying that the joint Bayesian method with $\delta \neq 0$ comes closer to capturing the true treatment assignment mechanism. As a result, estimates of β are similar in the joint Bayesian and traditional sequential estimation approaches, although the two methods do not produce the exact same estimates.

4.3 *Scenario where Every Covariate Exhibits Different Covariate-Treatment and Covariate-Outcome Relationships*

Finally, we simulate Scenario 3 so that every covariate exhibits different relationships with both the treatment and the outcome. For the PS model (11) we set $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6) = (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$. For the model in (12) we set $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = (0.0, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$.

We first analyze the data with $\delta = 0$. From Figure 3(a), we see that the joint Bayesian method provides estimates of $\gamma_1, \ldots, \gamma_6$ that are all shrunken toward 0.35 (the average value of $\gamma_1, \ldots, \gamma_6$), which is a consequence of estimating these quantities with feedback from an outcome model that assumes $(\gamma_1, \ldots, \gamma_6)$ can be rescaled by a single factor to fit the outcome-response surface, as discussed in Section 3.2. This is in stark contrast to the estimates from the sequential method that are not informed by the outcome and accurately reflect a different γ_k for $k = 1, 2, \ldots, 6$. We also see that these vast discrepancies between estimates of γ lead to estimates of β that are very different in the two methods, with the joint Bayesian estimates performing very poorly. Thus, in a setting where the covariate-treatment and covariate-outcome relationships are different for every covariate, joint Bayesian estimation with $\delta = 0$ cannot adequately recover the treatment effect, even though traditional sequential methods perform well.

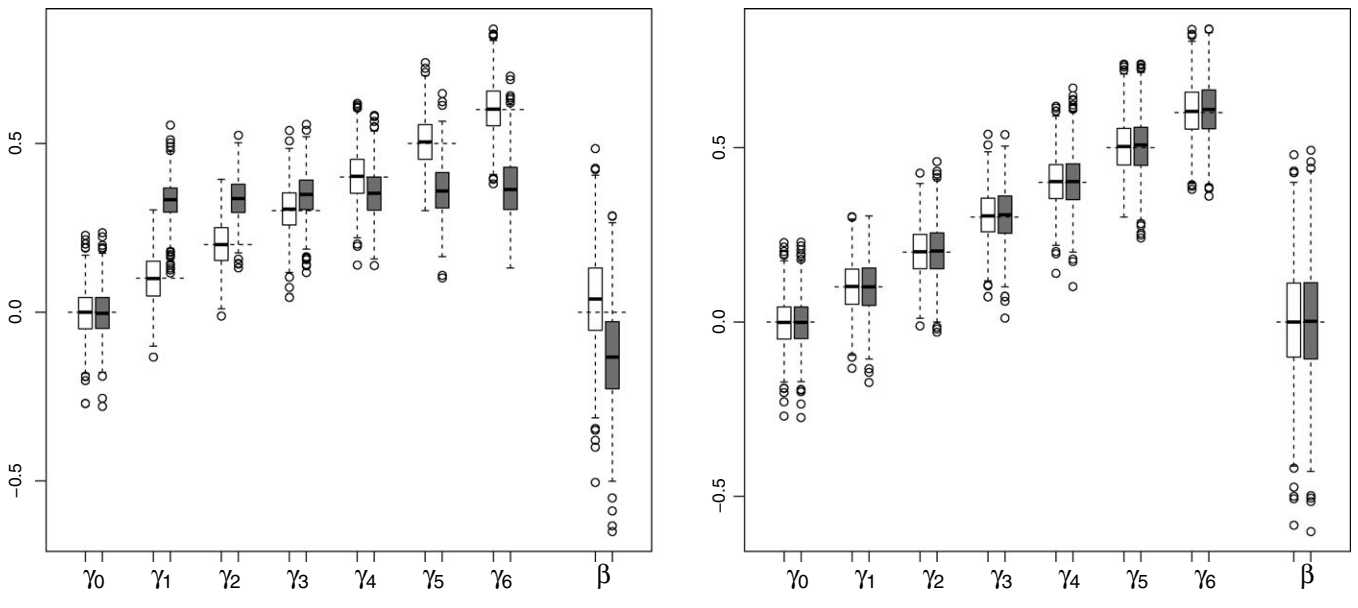
We reanalyze the data simulated in Scenario 3 with $\delta \neq 0$ and $C^+ = (C_1, \ldots, C_6)$, referring to this analysis as Scenario 3^+. Results for these analyses are summarized in Figure 3(b), which depicts that the additional covariate adjustment in the outcome model prevents feedback from distorting estimates of γ, leading to Bayesian estimates of γ that agree, on average, with those from the sequential procedure and the true treatment assignment mechanism. As a consequence, $h(\gamma, C)$ maintains the balancing-score property, and Bayesian estimates of β agree very closely with estimates from the sequential procedure and with the true parameter value.



(a) Scenario 2, $\delta = 0$

(b) Scenario $2^+, \delta \neq 0$

Figure 2. Scenarios 2 and 2^+ with $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6) = (0.0, 0.3, 0.3, 0.3, 0.3, 0.3, -0.3)$, and $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = (0.0, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$: boxplots of estimates of γ and β from the sequential frequentist and joint Bayesian analysis of 1000 replicated data sets. Shaded boxes are for the joint Bayesian analysis, unshaded are for the traditional sequential analysis. Horizontal dotted lines are at the true parameter values.



(a) Scenario 3, $\delta = 0$

(b) Scenario $3^+, \delta \neq 0$

Figure 3. Scenarios 3 and 3^+ with $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6) = (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$, and $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = (0.0, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$: boxplots of estimates of γ and β from the sequential frequentist and joint Bayesian analysis of 1000 replicated data sets. Shaded boxes are for the joint Bayesian analysis, unshaded are for the traditional sequential analysis. Horizontal dotted lines are at the true parameter values.

Table 1

Numerical performance comparison of estimates of β from the unadjusted analysis (Unadj), traditional sequential (Seq), joint Bayesian method (Bayes), and gold standard analysis (Gold). $\overline{\widehat{SE(\beta)}}$ is the average standard error estimates or posterior standard deviations across replicated data sets. Coverage refers to the proportion of 95% confidence or posterior probability intervals that contain β

	Bias($\widehat{\beta}$)				MSE($\widehat{\beta}$)			$\overline{\widehat{\text{SE}}(\widehat{\beta})}$			95 % Coverage		
	Unadj	Freq	Bayes	Gold	Freq	Bayes	Gold	Freq	Bayes	Gold	Freq	Bayes	Gold
Scenario 1	0.63	0.05	0.03	0	0.03	0.03	0.02	0.15	0.15	0.15	0.93	0.94	0.95
Scenario 2	0.42	0.04	−0.08	0	0.02	0.03	0.02	0.14	0.15	0.15	0.96	0.93	0.96
Scenario 2+	0.42	0.07	0.04	0	0.02	0.02	0.02	0.15	0.15	0.15	0.95	0.95	0.96
Scenario 3	0.4	0.04	−0.13	0	0.02	0.04	0.02	0.14	0.15	0.15	0.95	0.86	0.94
Scenario 3+	0.4	0	0	0	0.02	0.03	0.02	0.15	0.15	0.15	0.94	0.94	0.94

Table 1 numerically summarizes the performance of each method in terms of estimates of β. This table also summarizes bias from an unadjusted analysis, which is large in all scenarios. The joint Bayesian PS analyses of scenarios containing different covariate-outcome and covariate-treatment relationships that do not augment PS adjustment (Scenarios 2 and 3) produce estimates of β with substantial bias, as compared to the traditional sequential approach and to the Gold Standard analysis. Joint Bayesian estimates for these scenarios also exhibit low coverage probabilities.

For Scenarios 1, 2^+, and 3^+, estimates of β from the joint Bayesian and traditional sequential approach exhibit small bias. Posterior estimates are slightly more variable than the sequential estimates, as the Bayesian estimates reflect uncertainty in the PS. Coverage probabilities for estimates from the joint Bayesian and traditional sequential methods are comparable and near the nominal 95% rate for these scenarios. Note that the joint Bayesian estimates in Scenarios 1 and 2^+ exhibit a slight improvement in bias over the sequential procedure because these analyses jointly process the data while fixing certain components of δ exactly to zero in accordance with the known data-generating mechanism, but the researcher would not know which components of δ are exactly zero in practice.

It is also important to note that the detrimental effects of feedback on causal estimates when $\delta = 0$ (as displayed in Scenarios 2 and 3) is indeed a feature of the dimension-reduced feedback explicated in Section 3.2 and that this phenomenon cannot be remedied by increasingly flexible choices for $h(\gamma, C)$. To illustrate this point, the Web Appendix conducts a simulation study paralleling that in Scenarios 3 and 3^+, but specifying a flexible spline basis for $h(\gamma, C)$. The results of this simulation are the same as those presented here; posterior estimates of β perform poorly when $\delta = 0$, but not when $\delta \neq 0$, the latter case being analogous to the penalized spline of propensity prediction method of Little (2011).

5. Comparing the Effectiveness of Cardiovascular Treatments

CAS has recently emerged as a promising non-inferior alternative to CEA for treatment of carotid artery disease, which is a primary cause of stroke. To compare CEA ($X = 1$) vs. CAS ($X = 0$) for preventing death within one year of hospital admission ($Y = 1$ for death, 0 otherwise), we use hospital impatient data from 123,286 Medicare beneficiaries admitted to the hospital with a primary diagnosis of stroke during 2006 or 2007, as determined by the diagnosis codes found in Lichtman et al. (2009). An unadjusted comparison between 1-year mortality in CEA versus CAS patients yields an odds ratio for death of 0.58 indicating worse outcomes with CAS, but this comparison is thought to be confounded by patient characteristics that help determine treatment choice. In particular, CAS patients generally have a higher baseline risk profile, as evident from Table 2, which summarizes patient characteristics in the CEA and CAS treatment groups. In pursuit of a causal effect estimate, we conduct a PS analysis that adjusts for the 25 variables in Table 2, including patient ethnicity, age, and gender, as well as baseline risk factors consisting of the Hierarchical Condition Categories (Pope et al., 2004) for current or previous presence of comorbidities.

We conduct the analysis using logistic regression in both the PS and the outcome stages, with $h(\gamma, C)$ specifying PS subclasses based on quintiles of the logit(PS). We checked that the entire range of PS values was represented in both treatment groups (i.e., that there was sufficient overlap) using maximum likelihood estimates of γ. For the joint Bayesian method, the quintiles for defining PS subclasses were recalculated for every update of the PS. In light of the discussion in Sections 3 and 4, we consider an outcome model with $\delta \neq 0$ and $C^+ \equiv C$, implying residual adjustment for every covariate within PS subclass. We estimate the treatment effect using the joint Bayesian PS analysis of Section 3, as well as with a standard sequential analysis. Prior distributions for all parameters were considered Normal with mean 0 and variance 10^{10}. Three MCMC chains were run for 100,000 iterations, discarding the first 25,000 as burn in and saving every 20th sample for posterior inference. Note that convergence of the MCMC chains was adequate, but mixing was poor, requiring a large number of MCMC iterations.

5.1 *Results*

Figure 4 depicts how well each method balances covariates between CEA and CAS patients within PS subclass by depicting, for each binary covariate, the percent difference in prevalence between the treatment groups within each PS subclass. We see from Figure 4 that, with $\delta \neq 0$, joint Bayesian analysis and the traditional sequential analysis achieve similarly adequate covariate balance for the binary covariates. Age was also similarly balanced, with an average absolute

Table 2
Baseline characteristics (% experiencing unless noted) and 1-year mortality rate for CAS and CEA patients

	CAS (n=4038)	CEA (n=119,248)
Age (mean)	75.3	75.1
White	92.3	93.8
Male	62.1	57.3
Prior myocardial infarction	5.1	2.1
Unstable angina	5.2	2.5
Chronic atherosclerosis	64.3	48.6
Respiratory failure	3.3	1.9
Hypertension	75.3	78.8
Prior stroke	7.5	6.7
Cerebrovascular disease (nonstroke)	26.7	17.1
Renal failure	10.5	6
COPD	26.1	22.4
Pneumonia	5.4	3.6
Diabetes	35.3	32.3
Malnutrition	1.1	0.7
Dementia	3.6	3.1
Functional disability	5.1	3.8
Peripheral vascular disease	15.2	9
Trauma in the past year	4	3.4
Major psychiatric disorder	1	1
Anemia	15.5	12.3
Depression	3.9	4.7
Parkinsons /Huntingtons	1.1	0.8
Asthma	1.7	2.6
Cancer	4.7	4.2
Death within 1 year of admission	9.3	5.6

within-stratum difference between CEA and CAS patients of 0.18 years in both the joint Bayesian and traditional sequential analyses.

From the joint Bayesian PS analysis, the posterior mean of the conditional causal odds ratio (OR), e^{β}, was 0.68, with a 95% posterior probability interval (PI) of (0.61, 0.76), indicating a decreased odds of death within 1 year of hospital admission for CEA patients as compared to CAS patients. The analogous traditional sequential analysis produced the same point estimate and 95% confidence interval (CI). The posterior mean population causal OR was 0.69 (95% PI: 0.62, 0.77) with the joint Bayesian analysis, and the sequential procedure produced the same point estimate, with a 95% bootstrap CI of (0.63, 0.77) based on 1000 bootstrap samples. Thus, our analysis fails to provide evidence that CAS is a noninferior alternative to CEA for treating carotid artery disease in stroke patients, with increased odds of death within 1-year of hospital admission among patients treated with CAS. As in our simulation study, the joint Bayesian and traditional sequential analyses yield virtually identical results when $\delta \neq 0$.

Purely to illustrate the potential for feedback to distort the nature of the PS in an applied setting, we also conducted the joint Bayesian analysis that only adjusts for the PS in the outcome model ($\delta = 0$), even though our simulation study in Section 4 indicated that this procedure does not guarantee estimation of a causal effect. Poor estimation of the treatment-assignment mechanism when $\delta = 0$ is evident by the poor covariate balance relative to the analysis with $\delta \neq 0$ or the traditional sequential approach (see Figure 4). Estimates of the conditional causal OR were 0.69 (95% PI: 0.62, 0.78) with the joint Bayesian method with $\delta = 0$, and 0.67 (95% CI: 0.60,0.74) using the traditional sequential analysis with $\delta = 0$. We also note that for the Bayesian analysis with $\delta = 0$, MCMC performance was suspect for many parameters in the PS model, although convergence was adequate for all parameters in the outcome model, including β. We revisit this point in the discussion.

6. Discussion

Through a detailed assessment of model feedback, we have advanced the existing research on Bayesian PS estimation. Using a simple example and simulated illustrations, we have shown that a joint likelihood for a PS model and an outcome model that adjusts for the PS only cannot uncover treatment effects in general settings. The salient idea is that outcome models that adjust for the PS imply a characterization of the covariate-outcome response surface, and feedback from this outcome model can distort estimates from the PS stage and compromise the desirable features of PS adjustment. This casts substantial doubt on the validity of using Bayesian PS estimation for an outcome model that only adjusts for the PS, and represents a vital feature that has been previously overlooked in the literature on Bayesian PS estimation.

One constructive approach that we explore here augments PS adjustment with additional covariate adjustment, which has been previously recommended in the PS literature (Rubin, 1985; Stuart, 2010). We have shown that joint Bayesian PS estimation using this strategy can accurately estimate the treatment effect in settings where adjustment for only the PS fails. Our recommendation is that, when conducting joint Bayesian PS estimation with models for the PS and outcome stage, PS adjustment should be augmented with adjustment for every covariate that appears in the PS model, which is a strategy akin to those previously developed to yield "doubly robust" estimators that will estimate causal effects when either the PS model or the model for additional adjustment is correctly specified (Bang and Robins, 2005; Little, 2011). Although this strategy could still provide substantial benefit over methods for direct covariate adjustment that do not use the PS (Rubin, 1985), adjusting for each individual covariate within PS subclass may be unappealing to researchers drawn to PS methods precisely because of their ability to provide reliable causal estimates without specifying every covariate in an outcome model. If, when specifying a model for the PS and a model for the outcome, researchers wish not to augment PS adjustment with adjustment for every covariate, then we recommend against using the type of joint Bayesian PS estimation presented here.

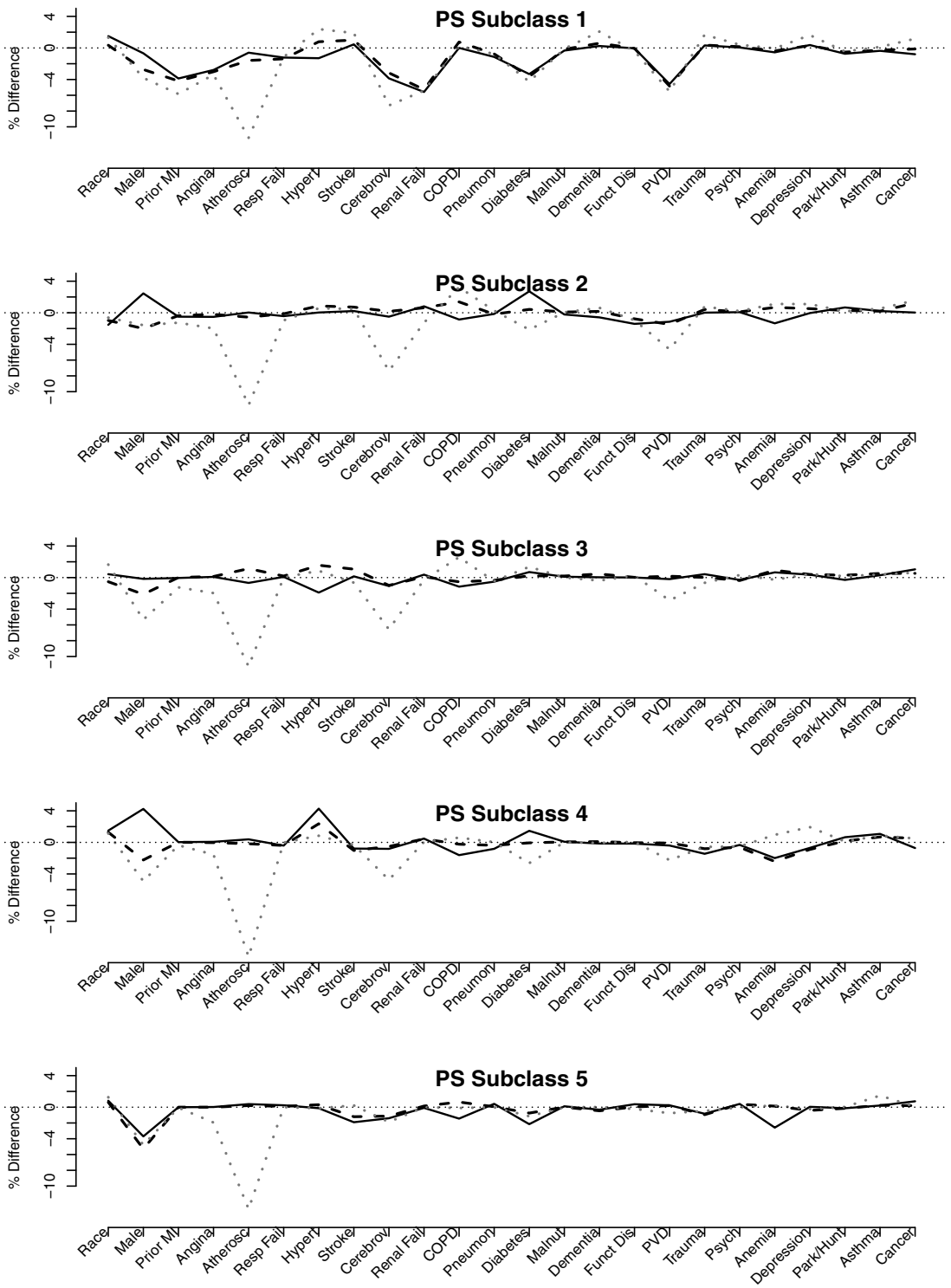


Figure 4. Summary of covariate balance within PS subclass for comparing the effectiveness of CEA versus CAS. Lines represent the difference in percent of $X = 0$ and $X = 1$ patients with each binary covariate within PS subclass. Solid line is for the traditional sequential analysis with $\delta \neq 0$, dashed line is for the joint Bayesian with $\delta \neq 0$, and the dotted line is for the joint Bayesian analysis with $\delta = 0$ (provided only for comparison). Balance from Joint Bayesian analyses is the posterior mean balance.

In comparison with traditional sequential procedures, joint Bayesian PS estimation implies a significant computational burden. In the analysis of the Medicare data, achieving adequate MCMC performance and chain mixing was challenging for parameters in the PS model, which can be considered nuisance parameters in a PS analysis. As noted in Section 5, MCMC performance for parameters in the PS model was poor in the analysis with $\delta = 0$, which was presented only for comparison. Regardless of the value of δ, performance of parameters in the outcome model, including β, was adequate.

Our goal for this work is to shed light on the subtlety of model feedback when conducting joint Bayesian PS estimation when a model is used to conduct outcome comparisons adjusted for the PS. To achieve this goal, we made several simplifying assumptions. In particular, we specified an outcome model that stratified on PS quintiles, but assumed the same treatment effect across all PS subclasses. In analyzing the Medicare stroke data, we investigated the use of additional PS subclasses and the inclusion of PS-by-treatment interaction terms in (2) to estimate a different treatment effect in each subclass, but this did not qualitatively alter our results. Other interactions or more complicated modeling strategies could be implemented in either the PS stage or the outcome stage, but the salient features of model feedback would persist, as shown the Web Appendix. We also note that the entire joint Bayesian PS estimation paradigm relies on a likelihood reflecting a PS model and an outcome model, and the issues addressed in this article have no clear analog to PS methods that exchange likelihood-based inference for matching or weighting in the outcome stage. Furthermore, the entirety of this article is predicated on the assumption of ignorable treatment assignment. Although this assumption held by design in our simulation study, our results regarding the comparative effectiveness of CEA vs. CAS should be viewed in light of the prospect of unmeasured confounding, which may be present in our example where the Medicare data lacks specific information on condition severity.

Better understanding of model feedback is essential to advance research on Bayesian methodology for use in problems involving the PS. For example, there has been recent interest in PS estimation when the set of necessary confounders is an unknown subset of those available for analysis (Wang, Parmigiani, and Dominici, 2012; McCandless, 2012; Vansteelandt, 2012). In principle, conducting Bayesian variable selection jointly on the PS and outcome models could ensure that important outcome predictors are included in the PS model, but our results here show that using model feedback to estimate coefficients in the PS model could prove detrimental. Although approximately Bayesian methods that "cut the feedback" (McCandless et al., 2010) could propagate uncertainty without distorting the PS, doing so in the context of variable selection would sacrifice the ability of the outcome to inform which variables to include in the PS. In another example relevant to joint Bayesian PS estimation, McCandless, Richardson, and Best (2012) use PS ideas to adjust for confounding using external validation data within a complex joint Bayesian model, but do not directly address the role of feedback. Investigation of feedback in these and other settings is an important avenue for future research, and provides sound motivation for further pursuit of Bayesian PS methods.

7. Supplementary Materials

The Web Appendix referenced in Sections 4 and 6 is available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

This work was funded by NCI P01CA134294, USEPA RD83479801, and HEI 4909. The contents of this work are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication. The authors thank Giovanni Parmigiani, Sebastien Haneuse, and Matt Cefalu for helpful discussion, as well as an Associate Editor and two referees for helpful comments.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88,** 669–679.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61,** 962–973.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Volume **648**: New York: Cambridge University Press.

Greenland, S., Robins, J., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science* **14,** 29–46.

Lichtman, J. H., Allen, N. B., Wang, Y., Watanabe, E., Jones, S. B., and Goldstein, L. B. (2009). Stroke patient outcomes in US hospitals before the start of the joint commission primary stroke center certification program. *Stroke* **40,** 3574–3579.

Little, R. (2011). Calibrated bayes, for statistics in general, and missing data in particular. *Statistical Science* **26,** 162–174.

McCandless, L., Richardson, S., and Best, N. (2012). Adjustment for missing confounders using external validation data and propensity scores. *Journal of the American Statistical Association* **107,** 40–51.

McCandless, L. C. (2012). Discussion of adjustment uncertainty and propensity scores. *Biometrics* **68,** 678–680.

McCandless, L. C., Douglas, I. J., Evans, S. J., and Smeeth, L. (2010). Cutting feedback in bayesian regression adjustment for the propensity score. *The International Journal of Biostatistics* **6,** 1–22.

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine* **28,** 94–112.

Pope, G. C., Kautter, J., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Lezzoni, L. I., Ingber, M. J., Levy, J. M., and Robst, J. (2004). Risk adjustment of medicare capitation payments using the CMS-HCC model. *Health Care Financing Review* **25,** 119–141.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

Rubin, D. (1985). The use of propensity scores in applied bayesian inference. In *Bayesian Statistics*, Volume **2,** 463–472. Valencia, Spain: Elsevier Science Publishers and Valencia University Press.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine* **26,** 20–36.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics* **2,** 808–840.

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20,** 512–522.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* **25,** 1–21.

Vansteelandt, S. (2012). Discussions. *Biometrics* **68,** 675–678.

Wang, C., Parmigiani, G., and Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* **68,** 661–671.

Received February 2012. *Revised September* 2012.
Accepted September 2012.