PSYCHOMETRIKA
2012
DOI: 10.1007/s11336-012-9262-8

A TWO-STEP BAYESIAN APPROACH FOR PROPENSITY SCORE ANALYSIS: SIMULATIONS AND CASE STUDY

DAVID KAPLAN AND JIANSHEN CHEN

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY, UNIVERSITY OF WISCONSIN–MADISON

A two-step Bayesian propensity score approach is introduced that incorporates prior information in the propensity score equation and outcome equation without the problems associated with simultaneous Bayesian propensity score approaches. The corresponding variance estimators are also provided. The two-step Bayesian propensity score is provided for three methods of implementation: propensity score stratification, weighting, and optimal full matching. Three simulation studies and one case study are presented to elaborate the proposed two-step Bayesian propensity score approach. Results of the simulation studies reveal that greater precision in the propensity score equation yields better recovery of the frequentist-based treatment effect. A slight advantage is shown for the Bayesian approach in small samples. Results also reveal that greater precision around the wrong treatment effect can lead to seriously distorted results. However, greater precision around the correct treatment effect parameter yields quite good results, with slight improvement seen with greater precision in the propensity score equation. A comparison of coverage rates for the conventional frequentist approach and proposed Bayesian approach is also provided. The case study reveals that credible intervals are wider than frequentist confidence intervals when priors are non-informative.

Key words: propensity score analysis, Bayesian inference.

1. Introduction

It is well established that for the purposes of warranting causal claims, a major advantage of experimental studies over quasi-experimental or observational studies is that the probability of an individual being assigned to a treatment group is known in advance. However, when experimental studies are not feasible, attention turns to the design of observational studies (Rosenbaum, 2002). In observational studies "treatments" are naturally occurring and the selection of individuals into these groups is governed by highly non-random and often unobservable processes.

In order to warrant causal inferences in the setting of observational studies, individuals in treatment conditions should be matched as closely as possible on observed pre-treatment assignment variables. In the case of unobserved pre-treatment variables, we must assume strong ignorability of treatment assignment (Rosenbaum & Rubin, 1983). Take, as an example, the effect of full versus part-day kindergarten on reading competency, in which random assignment is not feasible. To warrant a claim that full-day kindergarten boosts reading competency, a researcher would need to find children in the part-day kindergarten group who are as similar as possible on characteristics that might lead to selecting full- or part-day kindergarten. These characteristics should have been measured (or hypothetically present) before the child's selection into full- or part-day kindergarten (e.g. parental socio-economic status). Various forms of pre-treatment equating are available (see e.g. Rässler, 2002; Rubin, 2006). For this paper, we focus our attention on propensity score analysis as a method for equating groups on the basis of pre-treatment variables that are putatively related to the probability of having been observed in one or the other of the treatment conditions.

Requests for reprints should be sent to David Kaplan, Department of Educational Psychology, University of Wisconsin–Madison, 1025 W. Johnson St., Madison, WI 53706, USA. E-mail: dkaplan@education.wisc.edu

Published online: 30 March 2012

Propensity score analysis (PSA) has been used in a variety of settings, such as economics, education, epidemiology, psychology, and sociology. For comprehensive reviews see e.g. Guo and Fraser (2010), Steiner and Cook (in press), and Thoemmes and Kim (2011). Historically, propensity score analysis has been implemented within the frequentist perspective of statistics. Within that perspective, a considerable amount of attention has been paid to the problem of estimating the variance of the treatment effect. For example, early work on this problem by Rubin and Thomas (1992a, 1992b, 1996) found that matching on the estimated propensity score resulted in smaller variance compared to matching on the population propensity score. Thus, matching on the estimated propensity score is more efficient than matching on the true propensity score.

Other estimators of the treatment effect variance using propensity score weighting or matching approaches have been provided by several authors (Hirano, Imbens, & Ridder, 2003; Lunceford & Davidian, 2004), but few real data applications of their proposed methods have been available due to the requirement of large sample size and the computational complexity of the variance estimator. A recent variance estimator of the treatment effect based on matching was developed by Abadie and Imbens (2006), with an adjusted version developed by Abadie and Imbens (2009). Also, a bootstrap approach has been used to estimate the treatment effect variance (Lechner, 2002; Austin & Mamdani, 2006). However, Abadie and Imbens (2008) have shown that bootstrap inferences for the matching estimator are generally not valid due to the extreme non-smoothness of nearest neighbor matching.

In addition to the literature on frequentist-based propensity score analysis, there also exists literature examining propensity score analysis from a Bayesian perspective. This perspective views parameters as random and naturally accounts for uncertainty in the propensity score through the specification of prior distributions on propensity score model parameters. The purpose of this paper is to develop a straightforward two-step Bayesian approach to propensity score analysis, provide treatment effect and variance estimators and examine its behavior in the context of propensity score stratification, weighting, and optimal full matching.

The organization of this paper is as follows. For purposes of completeness, we first provide the nomenclature of the potential outcomes framework of causal inference, as developed by Rubin (1974). Next we review the conventional frequentist theory of propensity score analysis, focusing on three methods by which the propensity score is implemented: (a) subclassification, (b) weighting, and (c) optimal full matching. This is then followed by a discussion of the Bayesian propensity score. We then present the design of three simulation studies and one real data case study, followed by the results. The paper closes with the implications of our findings for the practice of using propensity scores to provide covariate balance when estimating treatment effects in observational studies.

2. The Neyman–Rubin Model of Causal Inference: Notation and Definitions

For this paper, we follow the general notation of the Neyman–Rubin potential outcomes model of causal inference Neyman (1923) and Rubin (1974), see also Holland (1986). To begin, let T be a treatment indicator, conveying the receipt of a treatment. In the case of a binary treatment, $T = \{0, 1\}$. For individual i, $T_i = 1$ if that individual received the treatment, and $T_i = 0$ if the individual did not receive the treatment. The essential idea of the Neyman–Rubin potential outcomes framework is that causal inference resides at the individual level. That is, for individual i, the goal, ideally, would be to observe the individual under receipt of the treatment and under non-receipt of the treatment. More formally, the *potential outcomes* framework for causal inference can be expressed as

$$Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i}, \tag{1}$$

where Y_i is the observed outcome of interest for person i, Y_{1i} is the potential outcome for individual i when exposed to the treatment, and Y_{0i} is the potential outcome for individual i when not exposed to the treatment. However, as Holland (1986) points out, the potential outcomes framework has a serious problem—namely, it is rarely possible to observe the values of Y_0 and Y_1 on the same individual i, and therefore rarely possible to observe the effects of $T = 1$ and $T = 0$. Holland refers to this as the *fundamental problem of causal inference*.

The statistical solution to the *fundamental problem* offered by Holland (1986) is to make use of the population of individuals. In this case, two causal estimands may be of interest. The first is the *average treatment effect*, γ_{ATE} for the target population defined as

$$\gamma_{\text{ATE}} = E(Y_1) - E(Y_0). \tag{2}$$

Returning again to the full-day versus part-day kindergarten example, (2) compares the effect of full-day kindergarten for those exposed to it, i.e. $E(Y_1|T = 1)$, to the effect of not being exposed to full-day kindergarten and exposed to part-day kindergarten instead, that is, $E(Y_0|T = 0)$. However, as noted by Heckman (2005) in most cases, the policy or clinical question lies in comparing $E(Y_1|T = 1)$ to the counterfactual group $E(Y_0|T = 1)$. In this case, the causal estimand of interest is the *average treatment effect on the treated, ATT*, defined as

$$\gamma_{\text{ATT}} = E(Y_1|T = 1) - E(Y_0|T = 1). \tag{3}$$

The essential difference between the ATE and ATT estimands can be seen by noting that for the ATE estimand it is assumed that a unit is drawn from a target population and assigned either to the treatment group or the control group. By contrast, the ATT estimand assumes that an individual is assigned to the treatment group and the question concerns the outcome of that individual had he/she been assigned to the control group. For this paper, we will concentrate on estimates of the ATE.

3. The Propensity Score

It was noted above that an essential requirement for warranting causal inferences in observational studies is that the treatment and control groups should be matched on all relevant covariates. In the context of random assignment experiments, matching takes place (at least asymptotically) at the group level by virtue of the randomization mechanism. Moreover, under random assignment, the randomization probabilities are known. Again, herein lies the power of the randomized experiment—namely, treated and untreated participants are matched on all observed and unobserved characteristics. In observational studies, the assignment mechanism is unknown, and hence we must make due with matching on as many relevant observed covariates as we have available. Thus, while matching on observed covariates is, in principle, attainable, there likely remains differences between treated and untreated participants on unobserved characteristics.

Clearly, better balancing can be achieved with a larger number of covariates. However, it is also the case that matching becomes prohibitively difficult as the number of covariates increases. An approach to handling a large number of covariates is to develop a so-called *balancing score*. Balancing scores are used to adjust treatment groups on the observed covariates so as to make them more comparable. A balancing score $b(z)$ is a function of the covariates z such that the conditional distribution of z given the balancing score is identical for the treatment group and control group. As noted by Rosenbaum and Rubin (1983), the finest balancing score $b(z)$ is the vector of all observed covariates z. Indeed, one or more balancing scores are used in a typical ANCOVA. The coarsest balancing score is the *propensity score* $e(z)$ which is a many-one function of the covariates z (Rosenbaum & Rubin, 1983, pg. 42).

More formally, consider first the potential outcomes model in (1). Under this model, the probability that individual i receives the treatment can be expressed as

$$e_i = p(T = 1|Y_{1i}, Y_{0i}, z_i, u_i), \tag{4}$$

where u_i contain unobserved covariates. Notice that in an observational study, (Y_{0i}, Y_{1i}, u_i) are not observed. Thus, it is not possible to obtain the true propensity score. Instead, we estimate the propensity score based on covariates z. Specifically,

$$\hat{e}(z) = p(T = 1|z), \tag{5}$$

which is referred to as the *estimated propensity score*.

The estimated propensity score $\hat{e}(z)$ has many important properties. Perhaps the most important property is the *balancing* property, which states that those in $T = 1$ and $T = 0$ with the same $\hat{e}(z)$ will have the same distribution on the covariates z. Formally, the balancing property can be expressed as

$$p\big\{z|T = 1, \hat{e}(z)\big\} = p\big\{z|T = 0, \hat{e}(z)\big\}, \tag{6}$$

or equivalently as

$$T \perp z|\hat{e}(z). \tag{7}$$

4. Implementation of the Propensity Score

In this section we describe three common approaches to the implementation of the propensity score: (a) stratification on $\hat{e}(z)$, (b) propensity score weighting, and (c) optimal full matching.

4.1. Stratification on $\hat{e}(z)$

There are numerous ways in which one can balance treatment groups on the observed covariates (see Rosenbaum & Rubin, 1983). For example, one approach would be to form strata directly on the basis of the observed covariates. However, as the number of covariates increases, the number of strata increases as well—such that it is unlikely that any given stratum would have enough members of all treatment groups to allow for reliable comparisons. The approach advocated in this paper is based on work by Cochran (1968) and utilizes subclassification into five strata based on the estimated propensity score. Subclassification into five strata on continuous distributions such as the propensity score has been shown to remove approximately 90 % of the bias due to non-random selection effects (Rosenbaum & Rubin, 1983; see also Cochran, 1968). However, for stratification on the propensity score to achieve the desired effect, the assumption of no hidden biases must hold.

Assuming no hidden biases, Rosenbaum and Rubin (1983) proved that when units within strata are homogeneous with respect to $\hat{e}(z)$, then the treatment and control units in the same stratum will have the same distribution on z. Moreover, Rosenbaum and Rubin showed that instead of using all of the covariates in z, a certain degree of parsimony can be achieved by using the coarser propensity score $\hat{e}(z)$. Finally, Rosenbaum and Rubin showed that if there are no hidden biases, then units with the same value on a balancing score (e.g., the propensity score), but assigned to different treatments, will serve as each other's control in that the expected difference in the responses of the units is equal to the average treatment effect.

4.2. *Propensity Score Weighting*

Still another approach to implementing the propensity score is based on weighting. Specifically, propensity score weighting is based on the idea of Horvitz–Thompson sampling weights (Horvitz & Thompson, 1952), and is designed to weight the treatment and control group participants in terms of their propensity scores. The details of this approach can be found in Hirano and Imbens (2001), Hirano et al. (2003), and Rosenbaum (1987).

As before, let $\hat{e}(z)$ be the estimated propensity score, and let T indicate whether an individual is treated ($T = 1$) or not ($T = 0$). The weight used to estimate the ATE can be defined as

$$\omega_1 = \frac{T}{\hat{e}(z)} + \frac{1 - T}{1 - \hat{e}(z)}. \tag{8}$$

Note that when $T = 1$, $\omega_1 = 1/\hat{e}(z)$ and when $T = 0$, $\omega_1 = 1/[1 - \hat{e}(z)]$. Thus, this approach weights the treatment and control group up to their respective populations.[1]

4.3. *Optimal Full Matching on the Propensity Score*

The third common approach for implementing the propensity score is based on the idea of statistical matching (see e.g. Hansen, 2004; Hansen & Klopfer, 2006; Rässler, 2002; Rosenbaum, 1989). Following Rosenbaum (1989), consider the problem of matching a treated unit to a control unit on a vector of covariates. In observational studies, the number of control units typically exceeds the number of treated units. A *matched pair* is an ordered pair (i, j), with $1 \leq i \leq N$ and $1 \leq j \leq M$ denoting that the ith treated unit is matched with the jth control unit. As defined by Rosenbaum (1989), "*A complete matched pair* is a set $\Im$ of N disjoint matched pairs, that is, N matched pairs in which each treated unit appears once, and each control unit appears either once or not at all" (pg. 1024).

Rosenbaum suggests two aspects of a "good" match. The first aspect is based on the notion of close matching in terms of a distance measure on the vector of covariates—for example, nearest neighbor matching. Obtaining close matches becomes more difficult as the number of covariates increases. Another aspect of a good match is based on covariate balance, for example, obtained on the propensity score. If distributions on the propensity score within matched samples are similar, then there is presumed to be balanced matching on the covariates.

For this paper, we consider *optimal matching*: an improvement on so-called *greedy matching*. Greedy matching finds a control unit to be matched to the treatment unit on the basis of the distance between those units alone. One form of the greedy algorithm works sequentially, starting with a match of minimum distance, and then removes the control unit and treated unit from further consideration. Then, the algorithm begins again. It is important to point out that greedy matching does not revisit the match, and therefore does not attempt to provide the lowest overall "cost" for the match.

Optimal matching, in contrast, proceeds much the same way as greedy matching. However, rather than simply adding a match, and removing the control (and treatment) from further consideration, optimal matching might reconsider a match if the total distance across matches is less than if the algorithm proceeded. According to Rosenbaum (1989), optimal matching is as good and often better than greedy matching. Indeed, although greedy matching can sometimes provide a good answer, there is no guarantee that the answer will be tolerable—and often it can be quite bad.

For this paper, we use the *optimal full matching* algorithm discussed in Hansen and Klopfer (2006) and implemented in their R package *optmatch* (Hansen & Klopfer, 2006). In optimal full

[1]The weight used to obtain an estimate of the ATT can be written as $\omega_2 = T + (1 - T)\frac{\hat{e}(z)}{1-\hat{e}(z)}$. Thus, the controls are weighted to the full sample using $1/[1 - \hat{e}(z)]$ and then further weighted to the treatment group using $\hat{e}(z)$.

matching, for each possible pair of units in control and treatment group with estimated propensity score $\hat{e}(z_0)$ and $\hat{e}(z_1)$, respectively, the distance is calculated by the difference between the estimated propensity scores, that is, $|\hat{e}(z_1) - \hat{e}(z_0)|$. The optimal full matching results are achieved when the total distance of matched propensity scores between treatment group and control group approaches the minimum. We do not utilize a caliper in this paper so that every observed unit is allowed to be matched.

5. Bayesian Propensity Score Approaches

Having outlined the conventional implementation of the propensity score, we now turn to more recent Bayesian approaches to the calculation of the propensity score. A recent systematic review by Thoemmes and Kim (2011) does not discuss Bayesian approaches to propensity score analysis, and our review of the extant literature also reveals very few studies examining Bayesian approaches to propensity score analysis. However, an earlier paper by Rubin (1985) argued that because propensity scores are, in fact, randomization probabilities, these should be of great interest to the applied Bayesian analyst.

Rubin contextualizes his arguments within the notion of the *well-calibrated* Bayesian (Dawid, 1982). Dawid's (1982) concept of calibration stems from the notion of forecasting within the Bayesian perspective.[2] Dawid uses weather forecasting as an example. Given a weather forecaster's subjective probability regarding tomorrow's weather, ω, she is well calibrated if, over a large number of forecast sequences, $p = \omega$, where p is the actual proportion of correct forecasts.[3]

In the context of propensity score analysis, Rubin (1985) argues that under the assumption of strong ignorability and assuming that the propensity score $\hat{e}(z)$ is an adequate summary of the observed covariates z, then our applied Bayesian will be well calibrated. To be specific, because the propensity score is the coarsest covariate that can be obtained and retain strong ignorability, and because there will be more data dedicated to estimating the parameters of interest (as opposed to the parameters associated with each covariate), the more reliable the forecasts—i.e. the better calibrated our applied Bayesian will be.

Although Rubin (1985) provides a justification for why an applied Bayesian should be interested in propensity scores, his analysis does not address the actual estimation of the propensity score equation or the outcome equation from a Bayesian perspective. In a more recent paper, Hoshino (2008) argued that propensity score analysis has focused mostly on estimating the marginal treatment effect and that more complex methods are needed to handle more realistic problems. In response, Hoshino (2008) developed a quasi-Bayesian estimation method that can be used to handle more general problems—and in particular, latent variable models.

More recently, McCandless, Gustafson, and Austin (2009) argued that the failure to account for uncertainty in the propensity score can result in falsely precise estimates of treatment effects. However, adopting the Bayesian perspective that data and parameters are random, appropriate consideration of uncertainty in model parameters of the propensity score equation can lead to a more accurate variance estimate of the treatment effect. In fact, it may be possible in many circumstances to elicit priors on the covariates from previous research or expert opinion and, as such, have a means of comparing different propensity score models for the same problem and resolve model choice via Bayesian model selection measures such as the deviance information criterion (Spiegelhalter, Best, Carlin, & van der Linde, 2002).

The paper by McCandless et al. (2009) provides an approach to Bayesian propensity score analysis for observational data. Their approach involves treating the propensity score as a latent

[2] As Dawid (1982) points out, *calibration* is synonymous with *reliability*.

[3] The concept of calibration can be extended to credible interval forecasts (see Dawid, 1982).

variable and modeling the joint likelihood of propensity scores and responses simultaneously in one Bayesian analysis via an MCMC algorithm. From there, the marginal posterior probability of the treatment effect can be obtained that directly incorporates uncertainty in the propensity score. Using a simulation study and a case study, McCandless et al. (2009) found that weak associations between the covariates and the treatment led to greater uncertainty in the propensity score and that the Bayesian subclassification approach yields wider credible intervals.

Following the McCandless et al. (2009) study, An (2010) presented a Bayesian approach that jointly models both the propensity score equation and outcome equation in one step and implemented this approach for propensity score regression and matching. The Bayesian approach of An (2010) was found to perform better for small samples. Also, An (2010) showed that frequentist PSA using the estimated propensity score tends to overestimate the standard error and has a larger estimated standard error than the Bayesian PSA approach, which contradicts McCandless et al. (2009). The difference between the estimator utilized by McCandless et al. (2009) and that utilized by An (2010) may contribute to the discrepancy in their findings. Specifically, McCandless et al. (2009) examined the variance estimator of frequentist PSA without any adjustment for uncertainty of the estimated propensity scores. An (2010), in contrast, did not clearly indicate how his variance estimator was obtained and most likely used a variance estimator that accounted for uncertainty of group assignment (An, 2010, pg. 15); perhaps building on work by Abadie and Imbens (2006, 2009). The frequentist PSA approach investigated in this paper is the same as McCandless et al. (2009).

Of relevance to the McCandless et al. (2009) and An (2010) approaches to Bayesian propensity score analysis, Gelman, Carlin, Stern, and Rubin (2003) have argued that the propensity score should provide information only regarding study design and not regarding the treatment effect, as is the case with the Bayesian procedure utilized by McCandless et al. (2009) and An (2010). Steiner and Cook (in press) also point out that the treatment might affect covariates and thus propensity score estimation should be based on covariates measured before treatment assignment. We agree, and view the McCandless et al. (2009) and An (2010) approaches as conceptually problematic. In particular, a possible consequence of these joint modeling approaches is that the predictive distribution of propensity scores will be affected by the outcome y, which can lead to a different propensity score estimate than obtained if y is not used in the analysis. Thus, to address the problem of joint modeling, this paper outlines a two-step modeling method using the Bayesian propensity score model in the first step, followed by a Bayesian outcome model in the second step. To investigate the Bayesian propensity score approach further, our paper employs the optimal full matching algorithm combined with propensity score stratification and weighting methods for both simulation and case studies. We provide the estimator of the treatment effect and the variance for our two-step approach. For comparison purposes, we also examine An's (2010) "intermediate Bayesian approach" which does address the joint modeling problem by specifying a Bayesian propensity score model in the first step with a conventional ordinary least squares outcome model in the second step. However, An (2010) does not provide detail regarding the variance estimator for this approach and only examines this intermediate Bayesian approach in the context of single nearest neighbor matching. Our paper provides a detailed discussion of the treatment effect and variance estimators and also examines An's (2010) intermediate Bayesian approach under stratification, weighting and optimal full matching.

6. Design and Results of Simulation Studies

In this section we present the design and results of two detailed simulation studies of our proposed two-step approach to Bayesian propensity score analysis (BPSA) and compare it to the conventional propensity score analysis (PSA). Also, we fit the simple linear regression and

Bayesian simple regression without any propensity score adjustment for the purpose of comparison. Bayesian simple regression utilizes the *MCMCregress* procedure (Martin, Quinn, & Park, 2010) in R (R Development Core Team, 2011) to draw from the posterior distribution of parameters of the outcome model using Gibbs sampling (Geman & Geman, 1984).

For this paper, we focus on the estimate of the average treatment effect defined in (2), which we denote as γ. The outcome model is written as a simple linear regression model—namely,

$$y = \mu + \gamma T + \epsilon_1, \tag{9}$$

where y is the outcome, μ is the intercept, γ is the average treatment effect, and T is the treatment indicator. In addition, we assume $\epsilon_1 \sim N(0, \sigma_1^2 I_n)$ where I_n is the n-dimensional identity matrix. Non-informative uniform priors are used for γ in Bayesian simple regression and an inverse gamma prior is used for σ_1^2, with shape parameter and scale parameter both 0.001.

For both PSA and BPSA, two models are specified. The first is a propensity score model, specified as the following logit model:

$$\text{Log}\left(\frac{e(z)}{1-e(z)}\right) = \alpha + \beta' z, \tag{10}$$

where α is the intercept, β is the slope vector and z represents a design matrix of chosen covariates. For BPSA, we utilized the R package *MCMClogit* (Martin et al., 2010) to draw from the posterior distribution of α and β in the logit model using a random walk Metropolis algorithm. After estimating propensity scores under PSA or BPSA, we use the outcome model in the second step to estimate the treatment effect γ via stratification, weighting, and optimal full matching.

In the two simulation studies, the estimated average treatment effect $\hat{\gamma}$ and standard error $\hat{\sigma}$ of the frequentist method come from ordinary least squares regression (OLS). Specifically, for conventional PSA, propensity score stratification is conducted by forming quintiles on the propensity score, calculating the OLS treatment effect within stratum, and averaging over the strata to obtain the treatment effect. The standard error was calculated as $\sqrt{\sum_{s=1}^{5} \hat{\sigma}_s^2/5}$, where $\hat{\sigma}_s^2$ is the variance of the estimated treatment effect for stratum s, and in which there were five strata. Although the assumption of independent stratifying units may incur bias, Zanutto, Lu, and Hornik (2005) have pointed out that this unadjusted variance estimator is a reasonable approximation to the standard error under the independence assumption and has often been used by researchers (e.g. Benjamin, 2003; Larsen, 1999; Perkins, Tu, Underhill, Zhou, & Murray, 2000). Propensity score weighting is performed by fitting a weighted regression with $1/\hat{e}(z)$ and $1/(1-\hat{e}(z))$ as the weights for the treatment and control group, respectively. Propensity score matching utilizes the optimal full matching method proposed by Hansen and Klopfer (2006). A group indicator is produced by assigning the matched participants into the same group and this serves as a covariate in the linear regression outcome model. The same simulated data set is used for PSA and BPSA in consideration of fair comparison, where PSA is denoted PSA-1rep in the tables. To take into account the sampling variability, 1000 replications are also conducted for traditional PSA, denoted PSA-1000rep. The details of Bayesian approach are provided in the simulation studies below and a flowchart outlining the steps of the simulation studies is shown on the left hand side of Figure 4.

6.1. Simulation Study 1

The first simulation study examines the Bayesian propensity score model and OLS outcome model across different sample sizes, true treatment effects, priors, and PSA methods, on estimates of the treatment effect γ. Because we use a Bayesian approach for the propensity score model only, we refer to this approach as *BPSA-1*, which is the same as An's (2010) "intermediate Bayesian approach". Data are generated according to the following procedure:

1. Independently generate random variables z_1, z_2 and z_3 as three covariates under sample size $n = 100$ and $n = 250$, respectively, such as

$$z_1 \sim N(1, 1)$$
$$z_2 \sim Poisson(2)$$
$$z_3 \sim Bernoulli(0.5).$$

   These distributions are chosen to imitate different types of covariates in practice such as continuous variables, count data and dichotomous (e.g. agree/disagree) variables.
2. Obtain the true propensity scores by

$$e(z) = \frac{\exp(0.2z_1 + 0.3z_2 - 0.2z_3)}{1 + \exp(0.2z_1 + 0.3z_2 - 0.2z_3)}, \tag{11}$$

   that is, the propensity score generating model has true $\alpha = 0$ and true $\beta = (0.2, 0.3, -0.2)'$.
3. Calculate the treatment assignment vector T by comparing the propensity score $e_i(z)$ to a random variable U_i generated from the $Uniform(0, 1)$ distribution, where $i = 1, \ldots, n$. Assign $T_i = 1$ if $U_i \leq e_i(z)$, $T_i = 0$ otherwise.
4. Generate outcomes $y_1, \ldots, y_n$ using the model:

$$y = 0.4z_1 + 0.3z_2 + 0.2z_3 + \gamma T + \epsilon_3, \tag{12}$$

   where $\epsilon_3 \sim N(0, 0.1)$ and γ is the true treatment effect taking two different values 0.25 and 1.25.
5. Data $= \{(y_i, z_i, T_i),\ i = 1, \ldots, n;\ n = 100 \text{ or } 250\}$.
6. Replicate the above steps 1000 times for the PSA model only.

The parameters α and β in the Bayesian propensity score model are standard regression coefficients. We assume a non-informative uniform prior on the intercept α and independent normal priors on β_k's ($k = 1$, 2 and 3 for three covariates, respectively):

$$\beta_k \sim N\left(b_\beta, B_\beta^{-1}\right),$$

with b_β as the prior mean and B_β as the prior precision. In simulation study 1, b_β is set as 0 to imitate the case of having little information on the mean, the same as what McCandless et al. (2009) chose in their study. Furthermore, we examine different prior precisions at $B_\beta = 0$, $B_\beta = 1$, $B_\beta = 10$ and $B_\beta = 100$ to explore the relation between the choice of prior precisions and the treatment effect. Note that when a prior precision takes value 0, the R program actually implements a non-informative uniform prior.

The MCMC sampling of the Bayesian propensity score model has 10^4 iterations with a thinning interval of 10 after 1000 burn-in. For n observations, there are $m = 1000$ sets of propensity scores $\hat{e}_{ij}$ ($i = 1, \ldots, n$, $j = 1, \ldots, m$) calculated from the joint posterior distribution of α and β. A treatment effect estimate $\hat{\gamma}_j$ is obtained using the jth set of estimated propensity scores $\hat{e}_{1j}(z), \ldots, \hat{e}_{nj}(z)$ by the conventional stratification, weighting, and optimal full matching methods as in traditional PSA. In fact, this is identical to the conventional approach, except we use a Bayesian propensity score model in the first step. The final estimate of the treatment effect in *BPSA-1* is

$$\hat{\gamma} = \frac{\sum_{j=1}^{m} \hat{\gamma}_j}{m}. \tag{13}$$

To estimate the variance of the treatment effect in BPSA-1, we assume independence among $\hat{\gamma}_j$s and thus

$$Var(\hat{\gamma}) = \frac{\sum_{j=1}^{m} Var(\hat{\gamma}_j)}{m^2}. \tag{14}$$

Because $\hat{\gamma}_j$s are estimated by the same outcome model with $\hat{e_j}$ sampled from the same posterior distribution, $\hat{\gamma}_j$s have the same distribution. Thus, $Var(\hat{\gamma_1}) = Var(\hat{\gamma_2}) = \ldots = Var(\hat{\gamma_m})$. Therefore, we have the following:

$$Var(\hat{\gamma}) = \frac{Var(\hat{\gamma}_1)}{m}. \tag{15}$$

For notational simplicity, let η denote the vector of propensity score model parameters. Then $Var(\hat{\gamma}_1)$ can be obtained via the total variance formula

$$Var(\hat{\gamma}_1) = E\big\{Var(\hat{\gamma}_1 \mid \eta)\big\} + Var\big\{E(\hat{\gamma}_1 \mid \eta)\big\}, \tag{16}$$

where $E\{Var(\hat{\gamma}_1 \mid \eta)\}$ is estimated by the average of $\hat{\sigma}_j^2$s; the conventional variance estimate of $\hat{\gamma}_j$ without considering the uncertainty of propensity scores produced by the OLS outcome model; that is,

$$E\big\{Var(\hat{\gamma}_1 \mid \eta)\big\} = m^{-1}\sum_{j=1}^{m}\hat{\sigma}_j^2, \tag{17}$$

and $Var\{E(\hat{\gamma}_1 \mid \eta)\}$ is estimated by the sample variance of $\hat{\gamma}_j$s; that is,

$$Var\big\{E(\hat{\gamma}_1 \mid \eta)\big\} = (m-1)^{-1}\sum_{j=1}^{m}\left(\hat{\gamma}_j - m^{-1}\sum_{j=1}^{m}\hat{\gamma}_j\right)^2. \tag{18}$$

By inserting (17) and (18) into (16) and taking into account (15), we obtain the following expression for the variance of the estimated treatment effects within the two-step BPSA approach:

$$Var(\hat{\gamma}) = \frac{m^{-1}\sum_{j=1}^{m}\hat{\sigma}_j^2 + (m-1)^{-1}\sum_{j=1}^{m}(\hat{\gamma}_j - m^{-1}\sum_{j=1}^{m}\hat{\gamma}_j)^2}{m}. \tag{19}$$

The BPSA variance estimation expression in (19) captures two sources of variation in the treatment effects, the variation from responses $\hat{\sigma}_j^2$s and the variation from propensity scores $\sum_{j=1}^{m}(\hat{\gamma}_j - m^{-1}\sum_{j=1}^{m}\hat{\gamma}_j)^2$. This second source of variation is often ignored in conventional PSA applications. Thus, our approach provides a variance estimation expression for the treatment effect while maintaining separation between the propensity score model and the outcome model in the estimation process.

6.2. Results of Simulation Study 1

The estimated average treatment effects $\hat{\gamma}$ and standard errors (SE) are shown in Table 1. With respect to bias in treatment effect estimates, both BPSA-1 (An's intermediate approach) and PSA obtain better estimates than simple linear regression without any adjustment. Generally, when there is insufficient prior information with low precision ($B_\beta = 0$ or 1), BPSA-1 provides similar $\hat{\gamma}$s to the conventional approach, with BPSA-1 yielding larger standard errors and thus wider credible intervals. Specifically, when $N = 250$, the SEs estimated by BPSA-1 are consistently larger than the SEs estimated by conventional PSA (stratification 0.09 vs. 0.07, weighting 0.13 vs. 0.08, and optimal full matching 0.09 or 0.08 vs. 0.07, for $B_\beta = 0$ or 1, respectively). Similar patterns are found for sample size $N = 100$. When the precision B_β increases to 10 or 100, the treatment effect estimates $\hat{\gamma}$s are closer to the true γ compared to conventional PSA, with smaller SEs. Note that study 1 utilized a prior mean of zero for α and β to imitate the situation of having little prior information on the propensity score parameters. We find that very high prior precision on the propensity score parameters does not necessarily provide good estimates. Instead, a moderate prior precision is safer and desirable.

Table 1 also shows that greater precision in the propensity score equation yields better recovery of the frequentist-based treatment effect compared to traditional PSA and compared to no

TABLE 1.
The $\hat{\gamma}$s & S.E.s for conventional PSA and BPSA with different true γs, sample sizes and prior precisions in study 1.

First Step	Second Step	$N = 100$		$N = 250$	
		$\gamma = 0.25$	$\gamma = 1.25$	$\gamma = 0.25$	$\gamma = 1.25$
PSA-1rep	Stratification	0.30 (0.13)	1.30 (0.13)	0.29 (0.07)	1.29 (0.07)
	Weighting	0.31 (0.12)	1.31 (0.12)	0.28 (0.08)	1.28 (0.08)
	Matching	0.23 (0.11)	1.23 (0.11)	0.29 (0.07)	1.29 (0.07)
PSA-1000rep	Stratification	0.28 (0.10)	1.28 (0.10)	0.28 (0.06)	1.28 (0.06)
	Weighting	0.26 (0.12)	1.26 (0.12)	0.25 (0.08)	1.25 (0.08)
	Matching	0.26 (0.09)	1.26 (0.09)	0.25 (0.05)	1.25 (0.05)
BPSA $B_\beta = 0$	Stratification	0.32 (0.13)	1.32 (0.13)	0.32 (0.09)	1.32 (0.09)
	Weighting	0.31 (0.18)	1.31 (0.18)	0.27 (0.13)	1.27 (0.13)
	Matching	0.27 (0.13)	1.27 (0.13)	0.31 (0.09)	1.31 (0.09)
BPSA $B_\beta = 1$	Stratification	0.31 (0.13)	1.31 (0.13)	0.31 (0.09)	1.31 (0.09)
	Weighting	0.29 (0.17)	1.29 (0.17)	0.26 (0.13)	1.26 (0.13)
	Matching	0.26 (0.13)	1.26 (0.13)	0.30 (0.08)	1.30 (0.08)
BPSA $B_\beta = 10$	Stratification	0.29 (0.12)	1.29 (0.12)	0.29 (0.07)	1.29 (0.07)
	Weighting	0.27 (0.15)	1.27 (0.15)	0.27 (0.11)	1.27 (0.11)
	Matching	0.24 (0.12)	1.24 (0.12)	0.27 (0.07)	1.27 (0.07)
BPSA $B_\beta = 100$	Stratification	0.29 (0.11)	1.29 (0.11)	0.29 (0.06)	1.29 (0.06)
	Weighting	0.28 (0.13)	1.28 (0.13)	0.35 (0.09)	1.35 (0.09)
	Matching	0.25 (0.11)	1.25 (0.11)	0.27 (0.06)	1.27 (0.06)
No Adjustment	SLR-1rep	0.35 (0.12)	1.35 (0.12)	0.54 (0.09)	1.54 (0.09)
	SLR-1000rep	0.48 (0.13)	1.48 (0.13)	0.48 (0.08)	1.48 (0.08)
	Bayes SLR	0.35 (0.12)	1.35 (0.12)	0.54 (0.09)	1.54 (0.09)

Note. SLR represents simple linear regression.

adjustment, especially when the sample size is relatively small. For $N = 100$, BPSA-1 with prior precisions of 10 and 100 obtain better $\hat{\gamma}$'s than PSA via stratification, weighting, or matching methods. But when N increases to 250, there is less advantage to BPSA-1 because the frequentist method is able to achieve better estimates due to more information from the data. PSA with 1000 replications offers more precise estimates than PSA with one replication, as expected. As a byproduct, we found that the optimal full matching method shows better estimates than stratification and weighting when $N = 100$, ranging from 0.23 to 0.27 for $\gamma = 0.25$ and 1.23 to 1.27 for $\gamma = 1.25$. The weighting approach works better when $N = 250$ and the prior precision is low. The size of the true treatment effects does not affect the pattern of estimates for both PSA and BPSA-1.

6.3. Simulation Study 2

To illustrate our two-step Bayesian approach to propensity score analysis, we conduct a second simulation study with both a Bayesian propensity score model and Bayesian outcome model, in which uniform priors were compared to normal priors with varying precision. We refer to this two-step fully Bayesian model as *BPSA-2*. Also, the effects of different sample sizes, priors and true γ's on the treatment effect are studied.

In study 2, the data generating process is the same as study 1. In addition, a Bayesian outcome model equation is developed according to (9) using the *MCMCregress* function in R, which

replaces the regular OLS outcome model for stratification and optimal full matching in study 1. However, to the best of our knowledge, Bayesian weighted regression has not yet been developed in the propensity score literature, thus a Bayesian weighing method with a Bayesian outcome model is beyond the scope of this paper.

The treatment effect and variance estimates of BPSA-2 are calculated in a different way from BPSA-1 because here we specify two Bayesian models. Again, let η denote the parameters of the propensity score model. For each observation, there are $m = 1000$ estimated propensity scores based on η_i, and for each estimated propensity score there are $J = 1000$ treatment effect estimates $\gamma_j(\eta)$'s sampled from posterior distribution of γ $(i = 1, \ldots, m,\ j = 1, \ldots, J)$. We estimate the treatment effect by the posterior mean of γ,

$$E(\gamma \mid T, y, z) = E\big\{E(\gamma \mid \eta, T, y, z) \,\big|\, T, y, z\big\}, \tag{20}$$

where $E(\gamma \mid \eta, T, y, z)$ is the posterior mean of γ in the Bayesian outcome model and can be estimated by $J^{-1}\sum_{j=1}^{J}\gamma_j(\eta)$. Then the treatment effect estimate is,

$$E(\gamma \mid T, y, z) = E\left\{J^{-1}\sum_{j=1}^{J}\gamma_j(\eta) \,\middle|\, T, y, z\right\}, \tag{21}$$

where $E\{J^{-1}\sum_{j=1}^{J}\gamma_j(\eta) \mid T, y, z\}$ can be estimated using the posterior sample mean of η from the Bayesian propensity score model, that is,

$$E\left\{J^{-1}\sum_{j=1}^{J}\gamma_j(\eta) \,\middle|\, T, y, z\right\} = m^{-1}J^{-1}\sum_{i=1}^{m}\sum_{j=1}^{J}\gamma_j(\eta_i). \tag{22}$$

The posterior variance of γ can be expressed as

$$\begin{aligned} Var(\gamma \mid T, y, z) = {} & E\big\{Var(\gamma \mid \eta, T, y, z) \,\big|\, T, y, z\big\} \\ & + Var\big\{E(\gamma \mid \eta, T, y, z) \,\big|\, T, y, z\big\}. \end{aligned} \tag{23}$$

In the first part of the right hand of (23), $Var(\gamma \mid \eta, T, y, z)$ can be estimated by the posterior sample variance $\sigma^2_{\gamma(\eta)}$ of γ in the Bayesian outcome model,

$$\sigma^2_{\gamma(\eta)} = (J-1)^{-1}\sum_{j=1}^{J}\left[\left\{\gamma_j(\eta) - J^{-1}\sum_{j=1}^{J}\gamma_j(\eta)\right\}\right]^2.$$

Thus,

$$E\big\{Var(\gamma \mid \eta, T, y, z) \,\big|\, T, y, z\big\} = E\big\{\sigma^2_{\gamma(\eta)} \,\big|\, T, y, z\big\} = m^{-1}\sum_{i=1}^{m}\sigma^2_{\gamma}(\eta_i). \tag{24}$$

In the second part of the right hand side of (23), $E(\gamma \mid \eta, T, y, z)$ can be estimated by the posterior sample mean $\mu_{\gamma(\eta)}$ of γ, where $\mu_{\gamma(\eta)} = J^{-1}\sum_{j=1}^{J}\gamma_j(\eta)$. Thus,

$$\begin{aligned} Var\big\{E(\gamma \mid \eta, T, y, z) \,\big|\, T, y, z\big\} &= Var\{\mu_{\gamma(\eta)} \mid T, y, z\} \\ &= (m-1)^{-1}\sum_{i=1}^{m}\left\{\mu_{\gamma(\eta_i)} - m^{-1}\sum_{i=1}^{m}\mu_{\gamma(\eta_i)}\right\}^2. \end{aligned} \tag{25}$$

Therefore, the posterior variance estimate of γ in BPSA-2 is

$$Var(\gamma \mid T, y, z) = m^{-1}\sum_{i=1}^{m}\sigma^2_{\gamma}(\eta_i) + (m-1)^{-1}\sum_{i=1}^{m}\left\{\mu_{\gamma(\eta_i)} - m^{-1}\sum_{i=1}^{m}\mu_{\gamma(\eta_i)}\right\}^2. \tag{26}$$

We note that (26) incorporates two components of variation. The first component of variation is the average of the posterior variances across the posterior samples, and the second component represents the variance of the posterior means across the posterior samples. Therefore, our BPSA-2 variance estimator is fully Bayesian, whereas An's intermediate approach (BPSA-1) is not fully Bayesian insofar as the outcome equation in An's approach is frequentist. Thus, we provide a fully Bayesian variance estimator so that researchers can incorporate priors in the outcome equation as well as in the propensity score equation.

For the Bayesian propensity score model, independent normal priors are chosen for α and β_k's ($k = 1, 2$ and 3 for three covariates, respectively):

$$\alpha \sim N\left(b_\alpha, B_\alpha^{-1}\right)$$
$$\beta_k \sim N\left(b_\beta, B_\beta^{-1}\right),$$

where b_α and b_β are prior means, and B_α and B_β are prior precisions. In the Bayesian outcome model, we also assume independent normal priors on the intercept μ and the treatment effect γ:

$$\mu \sim N\left(b_\mu, B_\mu^{-1}\right)$$
$$\gamma \sim N\left(b_\gamma, B_\gamma^{-1}\right),$$

with b_μ and b_γ as the prior mean, and B_μ and B_γ as the prior precisions.

There are two different designs for study 2, labeled I and II, to examine the performance of BPSA-2 when there is either little prior information or correct prior information on the means, respectively. In design I, we consider the situation of having little prior information on the means b_α, b_β, b_μ, and b_γ and set them as 0. Note that $B_\alpha = B_\beta = 0, 1, 10, 100$ in the first step and $B_\mu = B_\gamma = 0, 1, 10, 100$ in the second step.

In design II, true parameter values from the generating models are used as the prior means to imitate the case of having very accurate prior information on the means. For the Bayesian propensity score model, $b_\alpha = 0$ and $b_\beta = (0.2, 0.3, -0.2)'$. For the Bayesian outcome model, since the intercept μ is related to the propensity scores calculated in the first step and, moreover, since it may be difficult to elicit prior information of the intercept, we let $b_\mu = 0$ and $B_\mu = 0$ to indicate vague information on μ. The prior mean of the treatment effect $b_\gamma = 0.25$ or 1.25. The precisions also increase at $B_\alpha = B_\beta = 0, 1, 10, 100$ in the first step and $B_\gamma = 0, 1, 10, 100$ in the second step.

6.4. Results of Simulation Study 2

Results of design I are presented in Tables 2, 3, 4 and 5, while results of design II are presented in Tables 6, 7, 8 and 9. Examining Table 2 for example, the left hand most column shows the precision levels for the Bayesian propensity score model. The next column denotes the precision levels for the Bayesian outcome model. Then, the estimated treatment effects and standard errors follow.

For design I with little prior information on the mean, we find that given the same prior precision B_β for the propensity score model, the higher the prior precision B_γ on the treatment effect, the farther the treatment effect estimate is from true γ. This result indicates that greater precision around the wrong prior for the treatment effect can lead to seriously distorted results. However, consistent with the results of study 1, when $N = 100$ and B_γ is low, the low precisions ($B_\beta = 0$ or 1) on the propensity score model provide similar $\hat{\gamma}$s, but wider credible interval than traditional PSA, while the high precisions ($B_\beta = 10$ or 100) on the propensity score equation offer better treatment estimates and more concentrated intervals than PSA.

TABLE 2.
The $\hat{\gamma}$s & S.E.s for conventional and Bayesian stratification and matching of design I with true $\gamma = 0.25$ in study 2.

First Step	Second Step	$N = 100$		$N = 250$	
		Stratification	Matching	Stratification	Matching
$B_\beta = 0$	$B_\gamma = 0$	0.32 (0.14)	0.27 (0.13)	0.32 (0.09)	0.31 (0.09)
	$B_\gamma = 1$	0.34 (0.13)	0.28 (0.13)	0.33 (0.09)	0.32 (0.08)
	$B_\gamma = 10$	0.39 (0.09)	0.31 (0.11)	0.40 (0.08)	0.39 (0.08)
	$B_\gamma = 100$	0.11 (0.05)	0.38 (0.08)	0.29 (0.04)	0.52 (0.06)
$B_\beta = 1$	$B_\gamma = 0$	0.31 (0.14)	0.26 (0.13)	0.31 (0.09)	0.30 (0.08)
	$B_\gamma = 1$	0.33 (0.13)	0.26 (0.13)	0.32 (0.09)	0.31 (0.08)
	$B_\gamma = 10$	0.38 (0.09)	0.30 (0.11)	0.39 (0.08)	0.38 (0.08)
	$B_\gamma = 100$	0.11 (0.05)	0.38 (0.08)	0.29 (0.04)	0.52 (0.06)
$B_\beta = 10$	$B_\gamma = 0$	0.29 (0.13)	0.24 (0.12)	0.29 (0.07)	0.27 (0.07)
	$B_\gamma = 1$	0.31 (0.12)	0.24 (0.12)	0.30 (0.07)	0.28 (0.07)
	$B_\gamma = 10$	0.37 (0.09)	0.29 (0.10)	0.36 (0.07)	0.34 (0.07)
	$B_\gamma = 100$	0.11 (0.05)	0.38 (0.07)	0.29 (0.04)	0.51 (0.06)
$B_\beta = 100$	$B_\gamma = 0$	0.28 (0.12)	0.25 (0.11)	0.29 (0.07)	0.27 (0.06)
	$B_\gamma = 1$	0.30 (0.11)	0.25 (0.11)	0.30 (0.06)	0.28 (0.06)
	$B_\gamma = 10$	0.37 (0.08)	0.29 (0.10)	0.35 (0.06)	0.32 (0.06)
	$B_\gamma = 100$	0.12 (0.05)	0.38 (0.07)	0.31 (0.04)	0.51 (0.06)
PSA-1rep		0.30 (0.13)	0.23 (0.11)	0.29 (0.07)	0.29 (0.07)
PSA-1000rep		0.28 (0.10)	0.26 (0.09)	0.28 (0.06)	0.25 (0.05)

TABLE 3.
Confidence or credible intervals for conventional and Bayesian stratification and matching of design I with true $\gamma = 0.25$ in study 2.

First Step	Second Step	$N = 100$		$N = 250$	
		Stratification	Matching	Stratification	Matching
$B_\beta = 0$	$B_\gamma = 0$	(0.05, 0.59)	(0.02, 0.52)	(0.14, 0.50)	(0.13, 0.49)
	$B_\gamma = 1$	(0.09, 0.59)	(0.03, 0.53)	(0.15, 0.51)	(0.16, 0.48)
	$B_\gamma = 10$	(0.21, 0.57)	(0.09, 0.53)	(0.24, 0.56)	(0.23, 0.55)
	$B_\gamma = 100$	(0.01, 0.21)	(0.22, 0.54)	(0.21, 0.37)	(0.40, 0.64)
$B_\beta = 1$	$B_\gamma = 0$	(0.04, 0.58)	(0.01, 0.51)	(0.13, 0.49)	(0.14, 0.46)
	$B_\gamma = 1$	(0.08, 0.58)	(0.01, 0.51)	(0.14, 0.50)	(0.15, 0.47)
	$B_\gamma = 10$	(0.20, 0.56)	(0.08, 0.52)	(0.23, 0.55)	(0.22, 0.54)
	$B_\gamma = 100$	(0.01, 0.21)	(0.22, 0.54)	(0.21, 0.37)	(0.40, 0.64)
$B_\beta = 10$	$B_\gamma = 0$	(0.04, 0.54)	(0.00, 0.48)	(0.15, 0.43)	(0.13, 0.41)
	$B_\gamma = 1$	(0.07, 0.55)	(0.00, 0.48)	(0.16, 0.44)	(0.14, 0.42)
	$B_\gamma = 10$	(0.19, 0.55)	(0.09, 0.49)	(0.22, 0.50)	(0.20, 0.48)
	$B_\gamma = 100$	(0.01, 0.21)	(0.24, 0.52)	(0.21, 0.37)	(0.39, 0.63)
$B_\beta = 100$	$B_\gamma = 0$	(0.04, 0.52)	(0.03, 0.47)	(0.15, 0.43)	(0.15, 0.39)
	$B_\gamma = 1$	(0.08, 0.52)	(0.03, 0.47)	(0.18, 0.42)	(0.16, 0.40)
	$B_\gamma = 10$	(0.21, 0.53)	(0.09, 0.49)	(0.23, 0.47)	(0.20, 0.44)
	$B_\gamma = 100$	(0.02, 0.22)	(0.24, 0.52)	(0.23, 0.39)	(0.39, 0.63)
PSA-1rep		(0.05, 0.55)	(0.01, 0.45)	(0.15, 0.43)	(0.15, 0.43)
PSA-1000rep		(0.08, 0.48)	(0.08, 0.44)	(0.16, 0.40)	(0.15, 0.35)

TABLE 4.
The $\hat{\gamma}$s & S.E.s for conventional and Bayesian stratification and matching of design I with true $\gamma = 1.25$ in study 2.

First Step	Second Step	$N = 100$		$N = 250$	
		Stratification	Matching	Stratification	Matching
$B_\beta = 0$	$B_\gamma = 0$	1.32 (0.14)	1.27 (0.13)	1.32 (0.09)	1.31 (0.09)
	$B_\gamma = 1$	1.27 (0.13)	1.27 (0.13)	1.31 (0.09)	1.32 (0.08)
	$B_\gamma = 10$	1.00 (0.12)	1.22 (0.11)	1.22 (0.06)	1.35 (0.07)
	$B_\gamma = 100$	0.08 (0.05)	0.62 (0.11)	0.27 (0.06)	1.15 (0.06)
$B_\beta = 1$	$B_\gamma = 0$	1.31 (0.14)	1.26 (0.13)	1.31 (0.09)	1.30 (0.08)
	$B_\gamma = 1$	1.26 (0.13)	1.25 (0.13)	1.30 (0.08)	1.31 (0.08)
	$B_\gamma = 10$	1.00 (0.12)	1.21 (0.11)	1.22 (0.06)	1.34 (0.07)
	$B_\gamma = 100$	0.08 (0.05)	0.62 (0.11)	0.27 (0.06)	1.15 (0.06)
$B_\beta = 10$	$B_\gamma = 0$	1.29 (0.13)	1.24 (0.12)	1.29 (0.07)	1.27 (0.07)
	$B_\gamma = 1$	1.25 (0.12)	1.23 (0.12)	1.28 (0.07)	1.28 (0.07)
	$B_\gamma = 10$	1.00 (0.12)	1.20 (0.10)	1.22 (0.06)	1.31 (0.06)
	$B_\gamma = 100$	0.08 (0.05)	0.62 (0.11)	0.30 (0.07)	1.15 (0.06)
$B_\beta = 100$	$B_\gamma = 0$	1.28 (0.12)	1.25 (0.11)	1.29 (0.07)	1.27 (0.06)
	$B_\gamma = 1$	1.25 (0.11)	1.25 (0.11)	1.29 (0.06)	1.27 (0.05)
	$B_\gamma = 10$	1.06 (0.13)	1.22 (0.10)	1.25 (0.05)	1.30 (0.05)
	$B_\gamma = 100$	0.08 (0.05)	0.62 (0.11)	0.34 (0.09)	1.16 (0.06)
PSA-1rep		1.30 (0.13)	1.23 (0.11)	1.29 (0.07)	1.29 (0.07)
PSA-1000rep		1.28 (0.10)	1.26 (0.09)	1.28 (0.06)	1.25 (0.05)

TABLE 5.
Confidence or credible intervals for conventional and Bayesian stratification and matching of design I with true $\gamma = 1.25$ in study 2.

First Step	Second Step	$N = 100$		$N = 250$	
		Stratification	Matching	Stratification	Matching
$B_\beta = 0$	$B_\gamma = 0$	(1.05, 1.59)	(1.02, 1.52)	(1.14, 1.50)	(1.13, 1.49)
	$B_\gamma = 1$	(1.02, 1.52)	(1.02, 1.52)	(1.13, 1.49)	(1.16, 1.48)
	$B_\gamma = 10$	(0.76, 1.24)	(1.00, 1.44)	(1.10, 1.34)	(1.21, 1.49)
	$B_\gamma = 100$	(−0.02, 0.18)	(0.40, 0.84)	(0.15, 0.39)	(1.03, 1.27)
$B_\beta = 1$	$B_\gamma = 0$	(1.03, 1.57)	(1.01, 1.51)	(1.13, 1.49)	(1.14, 1.46)
	$B_\gamma = 1$	(1.01, 1.51)	(1.00, 1.50)	(1.14, 1.46)	(1.15, 1.47)
	$B_\gamma = 10$	(0.76, 1.24)	(0.99, 1.43)	(1.10, 1.34)	(1.20, 1.48)
	$B_\gamma = 100$	(−0.02, 0.18)	(0.40, 0.84)	(0.15, 0.39)	(1.03, 1.27)
$B_\beta = 10$	$B_\gamma = 0$	(1.04, 1.54)	(1.00, 1.48)	(1.15, 1.43)	(1.13, 1.41)
	$B_\gamma = 1$	(1.01, 1.49)	(0.99, 1.47)	(1.14, 1.42)	(1.13, 1.42)
	$B_\gamma = 10$	(0.76, 1.24)	(1.00, 1.40)	(1.10, 1.34)	(1.19, 1.43)
	$B_\gamma = 100$	(−0.02, 0.18)	(0.40, 0.84)	(0.16, 0.44)	(1.03, 1.27)
$B_\beta = 100$	$B_\gamma = 0$	(1.04, 1.52)	(1.03, 1.47)	(1.15, 1.43)	(1.15, 1.39)
	$B_\gamma = 1$	(1.03, 1.47)	(1.03, 1.47)	(1.17, 1.41)	(1.17, 1.37)
	$B_\gamma = 10$	(0.81, 1.31)	(1.02, 1.42)	(1.15, 1.35)	(1.20, 1.40)
	$B_\gamma = 100$	(−0.02, 0.18)	(0.40, 0.84)	(0.16, 0.52)	(1.04, 1.28)
PSA-1rep		(1.05, 1.55)	(1.01, 1.45)	(1.15, 1.43)	(1.15, 1.43)
PSA-1000rep		(1.08, 1.48)	(1.08, 1.44)	(1.16, 1.40)	(1.15, 1.35)

TABLE 6.
The $\hat{\gamma}$s & S.E.s for conventional and Bayesian stratification and matching of design II with true $\gamma = 0.25$ in study 2.

First Step	Second Step	$N = 100$		$N = 250$	
		Stratification	Matching	Stratification	Matching
$B_\beta = 0$	$B_\gamma = 0$	0.32 (0.14)	0.27 (0.13)	0.32 (0.09)	0.31 (0.09)
	$B_\gamma = 1$	0.31 (0.13)	0.27 (0.13)	0.32 (0.09)	0.31 (0.09)
	$B_\gamma = 10$	0.29 (0.10)	0.27 (0.12)	0.30 (0.08)	0.31 (0.08)
	$B_\gamma = 100$	0.26 (0.04)	0.26 (0.08)	0.26 (0.04)	0.29 (0.07)
$B_\beta = 1$	$B_\gamma = 0$	0.30 (0.14)	0.25 (0.13)	0.31 (0.09)	0.30 (0.08)
	$B_\gamma = 1$	0.30 (0.13)	0.25 (0.13)	0.31 (0.09)	0.30 (0.08)
	$B_\gamma = 10$	0.28 (0.10)	0.25 (0.12)	0.30 (0.07)	0.30 (0.08)
	$B_\gamma = 100$	0.26 (0.04)	0.25 (0.08)	0.26 (0.04)	0.28 (0.06)
$B_\beta = 10$	$B_\gamma = 0$	0.28 (0.13)	0.23 (0.11)	0.28 (0.07)	0.27 (0.07)
	$B_\gamma = 1$	0.27 (0.12)	0.23 (0.11)	0.28 (0.07)	0.27 (0.07)
	$B_\gamma = 10$	0.26 (0.09)	0.23 (0.11)	0.27 (0.06)	0.27 (0.07)
	$B_\gamma = 100$	0.25 (0.04)	0.24 (0.07)	0.25 (0.04)	0.26 (0.06)
$B_\beta = 100$	$B_\gamma = 0$	0.24 (0.09)	0.23 (0.09)	0.26 (0.06)	0.24 (0.05)
	$B_\gamma = 1$	0.24 (0.09)	0.23 (0.09)	0.26 (0.06)	0.24 (0.05)
	$B_\gamma = 10$	0.25 (0.08)	0.23 (0.08)	0.25 (0.06)	0.24 (0.05)
	$B_\gamma = 100$	0.25 (0.04)	0.24 (0.06)	0.24 (0.04)	0.24 (0.05)
PSA-1rep		0.30 (0.13)	0.23 (0.11)	0.29 (0.07)	0.29 (0.07)
PSA-1000rep		0.28 (0.10)	0.26 (0.09)	0.28 (0.06)	0.25 (0.05)

TABLE 7.
Confidence interval or credible interval for conventional and Bayesian stratification and matching of design II with true $\gamma = 0.25$ in study 2.

First Step	Second Step	$N = 100$		$N = 250$	
		Stratification	Matching	Stratification	Matching
$B_\beta = 0$	$B_\gamma = 0$	(0.05, 0.59)	(0.02, 0.52)	(0.14, 0.50)	(0.13, 0.49)
	$B_\gamma = 1$	(0.06, 0.56)	(0.02, 0.52)	(0.14, 0.50)	(0.13, 0.49)
	$B_\gamma = 10$	(0.09, 0.49)	(0.03, 0.51)	(0.14, 0.46)	(0.15, 0.47)
	$B_\gamma = 100$	(0.18, 0.34)	(0.10, 0.42)	(0.18, 0.34)	(0.15, 0.43)
$B_\beta = 1$	$B_\gamma = 0$	(0.03, 0.57)	(0.00, 0.50)	(0.13, 0.49)	(0.14, 0.46)
	$B_\gamma = 1$	(0.05, 0.55)	(0.00, 0.50)	(0.13, 0.49)	(0.14, 0.46)
	$B_\gamma = 10$	(0.08, 0.48)	(0.01, 0.49)	(0.16, 0.44)	(0.14, 0.46)
	$B_\gamma = 100$	(0.18, 0.34)	(0.09, 0.41)	(0.18, 0.34)	(0.16, 0.40)
$B_\beta = 10$	$B_\gamma = 0$	(0.03, 0.53)	(0.01, 0.45)	(0.14, 0.42)	(0.13, 0.41)
	$B_\gamma = 1$	(0.03, 0.51)	(0.01, 0.45)	(0.14, 0.42)	(0.13, 0.41)
	$B_\gamma = 10$	(0.08, 0.44)	(0.01, 0.45)	(0.15, 0.39)	(0.14, 0.41)
	$B_\gamma = 100$	(0.17, 0.33)	(0.10, 0.38)	(0.17, 0.33)	(0.14, 0.38)
$B_\beta = 100$	$B_\gamma = 0$	(0.06, 0.42)	(0.05, 0.41)	(0.14, 0.38)	(0.14, 0.34)
	$B_\gamma = 1$	(0.06, 0.42)	(0.05, 0.41)	(0.14, 0.38)	(0.14, 0.34)
	$B_\gamma = 10$	(0.09, 0.41)	(0.07, 0.39)	(0.13, 0.37)	(0.14, 0.34)
	$B_\gamma = 100$	(0.17, 0.33)	(0.12, 0.36)	(0.16, 0.32)	(0.14, 0.34)
PSA-1rep		(0.05, 0.55)	(0.01, 0.45)	(0.15, 0.43)	(0.15, 0.43)
PSA-1000rep		(0.08, 0.48)	(0.08, 0.44)	(0.16, 0.40)	(0.15, 0.35)

TABLE 8.
The $\hat{\gamma}$s & S.E.s for conventional and Bayesian stratification and matching of design II with true $\gamma = 1.25$ in study 2.

First Step	Second Step	$N = 100$		$N = 250$	
		Stratification	Matching	Stratification	Matching
$B_\beta = 0$	$B_\gamma = 0$	1.32 (0.14)	1.27 (0.13)	1.32 (0.09)	1.31 (0.09)
	$B_\gamma = 1$	1.31 (0.13)	1.27 (0.13)	1.32 (0.09)	1.31 (0.09)
	$B_\gamma = 10$	1.29 (0.10)	1.27 (0.12)	1.30 (0.08)	1.31 (0.08)
	$B_\gamma = 100$	1.26 (0.04)	1.26 (0.08)	1.26 (0.04)	1.29 (0.07)
$B_\beta = 1$	$B_\gamma = 0$	1.30 (0.14)	1.25 (0.13)	1.31 (0.09)	1.30 (0.08)
	$B_\gamma = 1$	1.30 (0.13)	1.25 (0.13)	1.31 (0.09)	1.30 (0.08)
	$B_\gamma = 10$	1.28 (0.10)	1.25 (0.12)	1.30 (0.07)	1.30 (0.08)
	$B_\gamma = 100$	1.26 (0.04)	1.25 (0.08)	1.26 (0.04)	1.28 (0.06)
$B_\beta = 10$	$B_\gamma = 0$	1.28 (0.13)	1.23 (0.11)	1.28 (0.07)	1.27 (0.07)
	$B_\gamma = 1$	1.27 (0.12)	1.23 (0.11)	1.28 (0.07)	1.27 (0.07)
	$B_\gamma = 10$	1.26 (0.09)	1.23 (0.11)	1.27 (0.06)	1.27 (0.07)
	$B_\gamma = 100$	1.25 (0.04)	1.24 (0.07)	1.25 (0.04)	1.26 (0.06)
$B_\beta = 100$	$B_\gamma = 0$	1.24 (0.09)	1.23 (0.09)	1.26 (0.06)	1.24 (0.05)
	$B_\gamma = 1$	1.24 (0.09)	1.23 (0.09)	1.26 (0.06)	1.24 (0.05)
	$B_\gamma = 10$	1.25 (0.08)	1.23 (0.08)	1.25 (0.06)	1.24 (0.05)
	$B_\gamma = 100$	1.25 (0.04)	1.24 (0.06)	1.24 (0.04)	1.24 (0.05)
PSA-1rep		1.30 (0.13)	1.23 (0.11)	1.29 (0.07)	1.29 (0.07)
PSA-1000rep		1.28 (0.10)	1.26 (0.09)	1.28 (0.06)	1.25 (0.05)

TABLE 9.
Confidence interval or credible interval for conventional and Bayesian stratification and matching of design II with true $\gamma = 1.25$ in study 2.

First Step	Second Step	$N = 100$		$N = 250$	
		Stratification	Matching	Stratification	Matching
$B_\beta = 0$	$B_\gamma = 0$	(1.05, 1.59)	(1.02, 1.52)	(1.14, 1.50)	(1.13, 1.49)
	$B_\gamma = 1$	(1.06, 1.56)	(1.02, 1.52)	(1.14, 1.50)	(1.13, 1.49)
	$B_\gamma = 10$	(1.09, 1.49)	(1.03, 1.51)	(1.14, 1.46)	(1.15, 1.47)
	$B_\gamma = 100$	(1.18, 1.34)	(1.10, 1.42)	(1.18, 1.34)	(1.15, 1.43)
$B_\beta = 1$	$B_\gamma = 0$	(1.03, 1.57)	(1.00, 1.50)	(1.13, 1.49)	(1.14, 1.46)
	$B_\gamma = 1$	(1.05, 1.55)	(1.00, 1.50)	(1.13, 1.49)	(1.14, 1.46)
	$B_\gamma = 10$	(1.08, 1.48)	(1.01, 1.49)	(1.16, 1.44)	(1.14, 1.46)
	$B_\gamma = 100$	(1.18, 1.34)	(1.09, 1.41)	(1.18, 1.34)	(1.16, 1.40)
$B_\beta = 10$	$B_\gamma = 0$	(1.03, 1.53)	(1.01, 1.45)	(1.14, 1.42)	(1.13, 1.41)
	$B_\gamma = 1$	(1.03, 1.51)	(1.01, 1.45)	(1.14, 1.42)	(1.13, 1.41)
	$B_\gamma = 10$	(1.08, 1.44)	(1.01, 1.45)	(1.15, 1.39)	(1.14, 1.41)
	$B_\gamma = 100$	(1.17, 1.33)	(1.10, 1.38)	(1.17, 1.33)	(1.14, 1.38)
$B_\beta = 100$	$B_\gamma = 0$	(1.06, 1.42)	(1.05, 1.41)	(1.14, 1.38)	(1.14, 1.34)
	$B_\gamma = 1$	(1.06, 1.42)	(1.05, 1.41)	(1.14, 1.38)	(1.14, 1.34)
	$B_\gamma = 10$	(1.09, 1.41)	(1.07, 1.39)	(1.13, 1.37)	(1.14, 1.34)
	$B_\gamma = 100$	(1.17, 1.33)	(1.12, 1.36)	(1.16, 1.32)	(1.14, 1.34)
PSA-1rep		(1.05, 1.55)	(1.01, 1.45)	(1.15, 1.43)	(1.15, 1.43)
PSA-1000rep		(1.08, 1.48)	(1.08, 1.44)	(1.16, 1.40)	(1.15, 1.35)

We find that BPSA-2 still performs better for $N = 100$. When N increases to 250, the treatment effect estimates of BPSA-2 and PSA approach each other, as expected. Also, the optimal full matching method performs slightly better than propensity score stratification. Different true γs yield similar patterns of treatment effect estimates.

For design II with true parameters as prior means, conditional on the same prior precision B_β for the propensity score model, greater prior precision on B_γ yields more accurate treatment effect estimates. In fact, according to Table 6 and Table 8, when $N = 100$ and $B_\gamma = 100$, the estimates are very close to the true γ, ranging from 0.24 to 0.26 for $\gamma = 0.25$ and 1.24 to 1.26 for $\gamma = 1.25$. These results suggest that greater precision around the correct treatment effect parameter yields quite ideal results.

Given the same prior precision B_γ on the outcome model, there is a slight improvement seen with greater precision in the propensity score equation, except for the matching method when $N = 100$. The violation may be due to the non-informative prior on the intercept of outcome model, but note that although estimates via matching do not hold the pattern for $N = 100$, the results are still good and stable, ranging from 0.23 to 0.27 for $\gamma = 0.25$ and 1.23 to 1.27 for $\gamma = 1.25$.

In contrast to design I, both stratification and optimal full matching work well when precisions are relatively high. When both B_β and B_γ are 100, stratification performs slightly better than optimal full matching.

In addition to the tabled results, we provide selected posterior density plots of BPSA-1 and BPSA-2 for a randomly selected subject displayed in Figures 1–3.[4] Figure 1 shows the posterior distribution of intercept α and slopes β_1, β_2 and β_3 of three covariates in the Bayesian propensity score equation, which suggests again that the Bayesian propensity score model naturally takes into account the uncertainty of propensity score estimates. The left column of Figure 1 represents the Bayesian propensity score model with non-informative uniform priors ($B_\beta = 0$), whereas the right column represents the Bayesian propensity score model with informative independent normal priors (that is, true prior means with prior precisions $B_\beta = 10$). The posterior distribution of parameters with informative priors is not only more concentrated, as expected, but also smoother than the posterior distribution with non-informative priors.

On the basis of the Bayesian propensity score estimates in Figure 1, the OLS outcome model and Bayesian outcome model estimates are shown in Figure 2 and Figure 3, respectively. Figure 2 illustrates the predictive distribution of treatment effect estimates $\hat{\gamma}$ by OLS regression based on the Bayesian propensity score estimates (BPSA-1). The left column of Figure 2 refers to stratification, weighting and optimal full matching estimates based on propensity score estimates with non-informative priors, while the right column indicates the distribution of $\hat{\gamma}$ corresponding to Bayesian propensity score estimates with informative priors. Weighting and optimal full matching approaches show more accurate estimates than the stratification method (True $\gamma = 1.25$). Figure 3 shows the distribution of the posterior means of $\hat{\gamma}$ estimated by the Bayesian outcome model (BPSA-2). The left and right columns represent stratification and optimal full matching estimates, respectively, while different rows imply different kinds of prior. The first row of Figure 3 indicates that the propensity score model and outcome model have non-informative priors; the second row refers to the non-informative propensity score model together with an informative outcome model, the third row is the opposite of the second row and has an informative propensity score model only, and the fourth row has informative priors for both models. According to Figure 3, informative priors on the outcome model yield more accurate treatment effect estimates (second and fourth row in Figure 3) regardless of whether informative or non-informative priors are used in the propensity score model. This suggests that accurate prior information on the treatment effect has a greater impact on the estimate than prior information on the propensity

[4] All other density plots and convergence plots are available upon request.

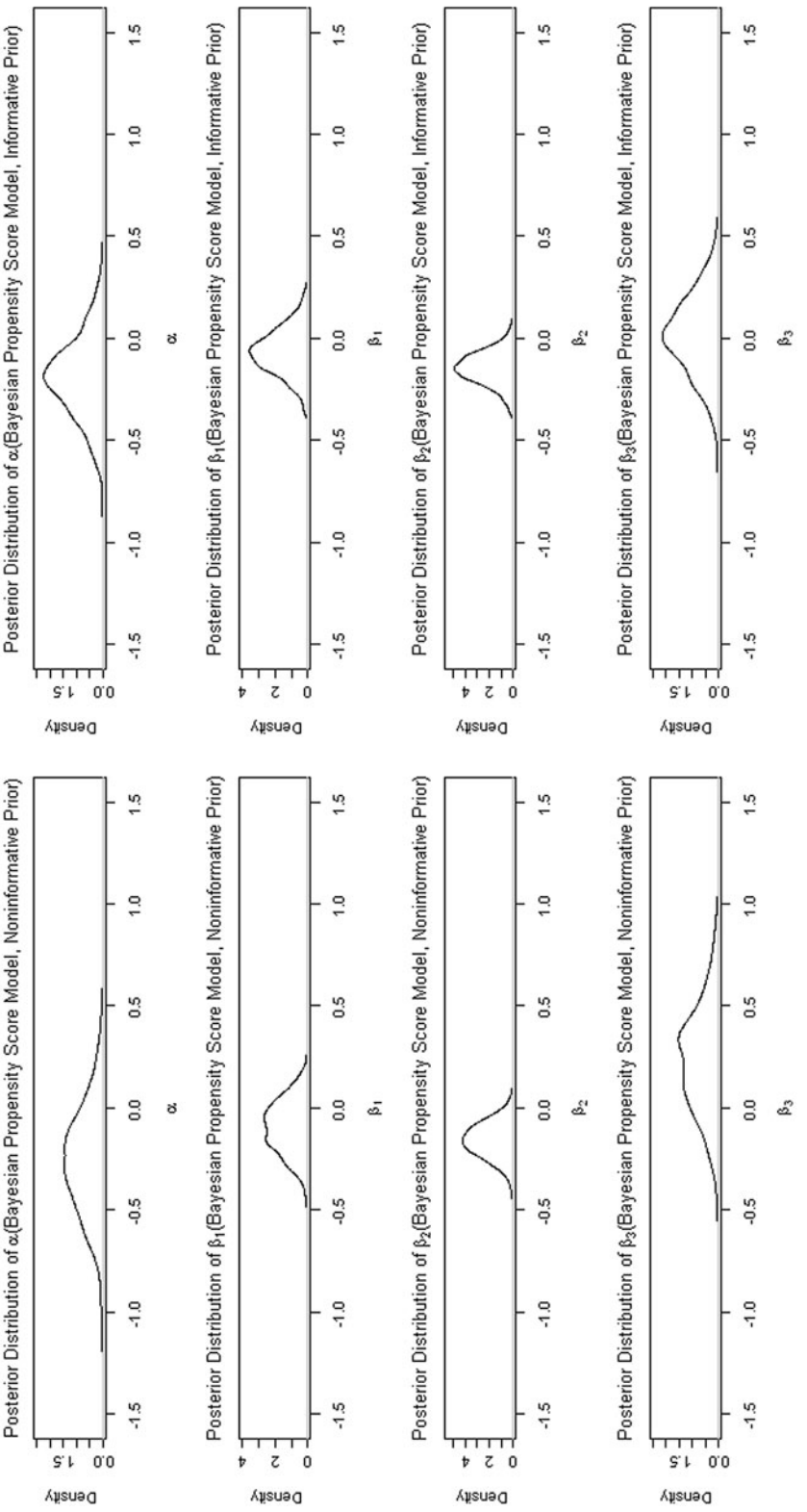

FIGURE 1.
Posterior density plots of parameters in Bayesian propensity score equation.

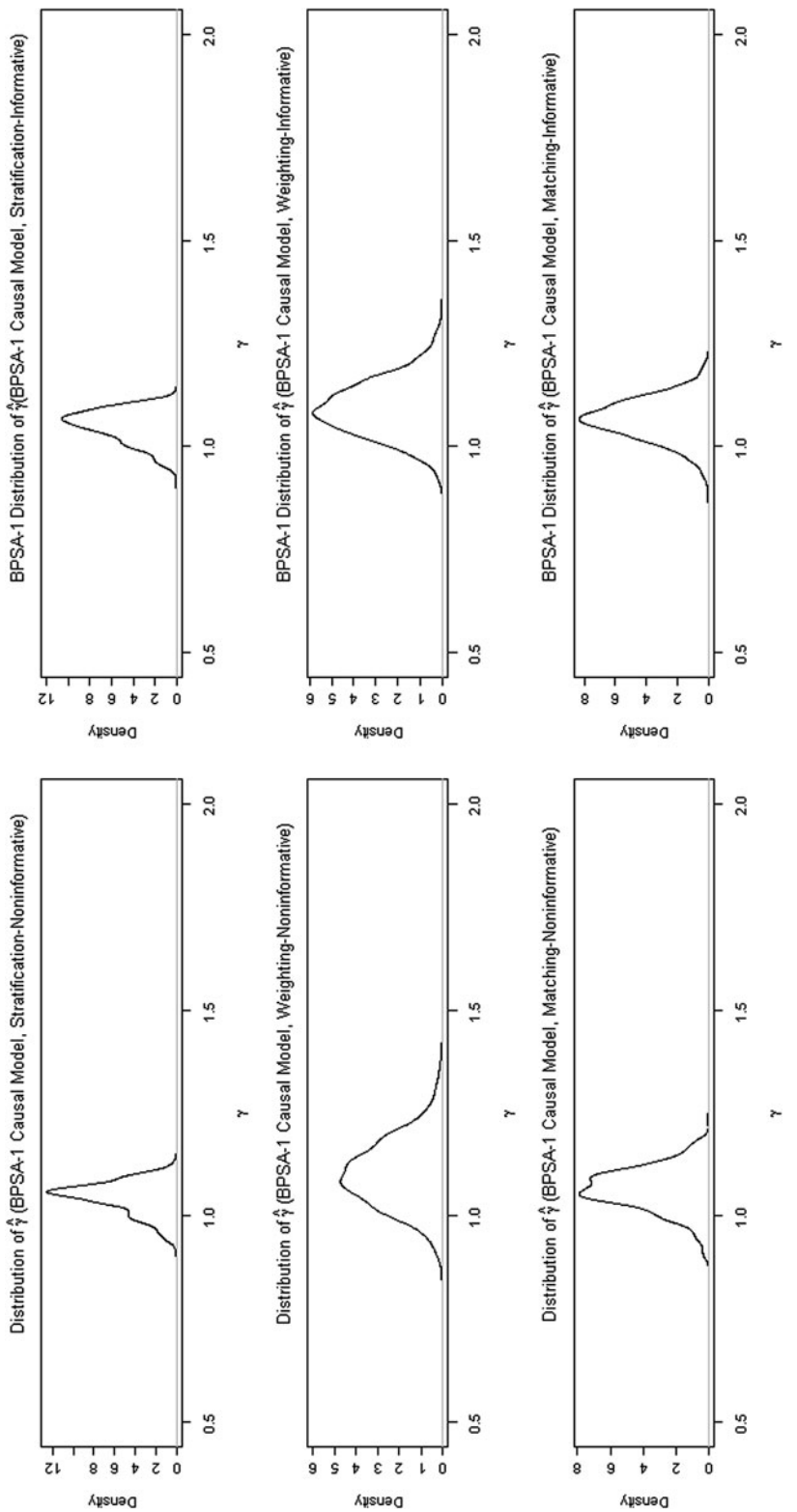

FIGURE 2.
Distribution of $\hat{\gamma}$ in BPSA-1 outcome equation.

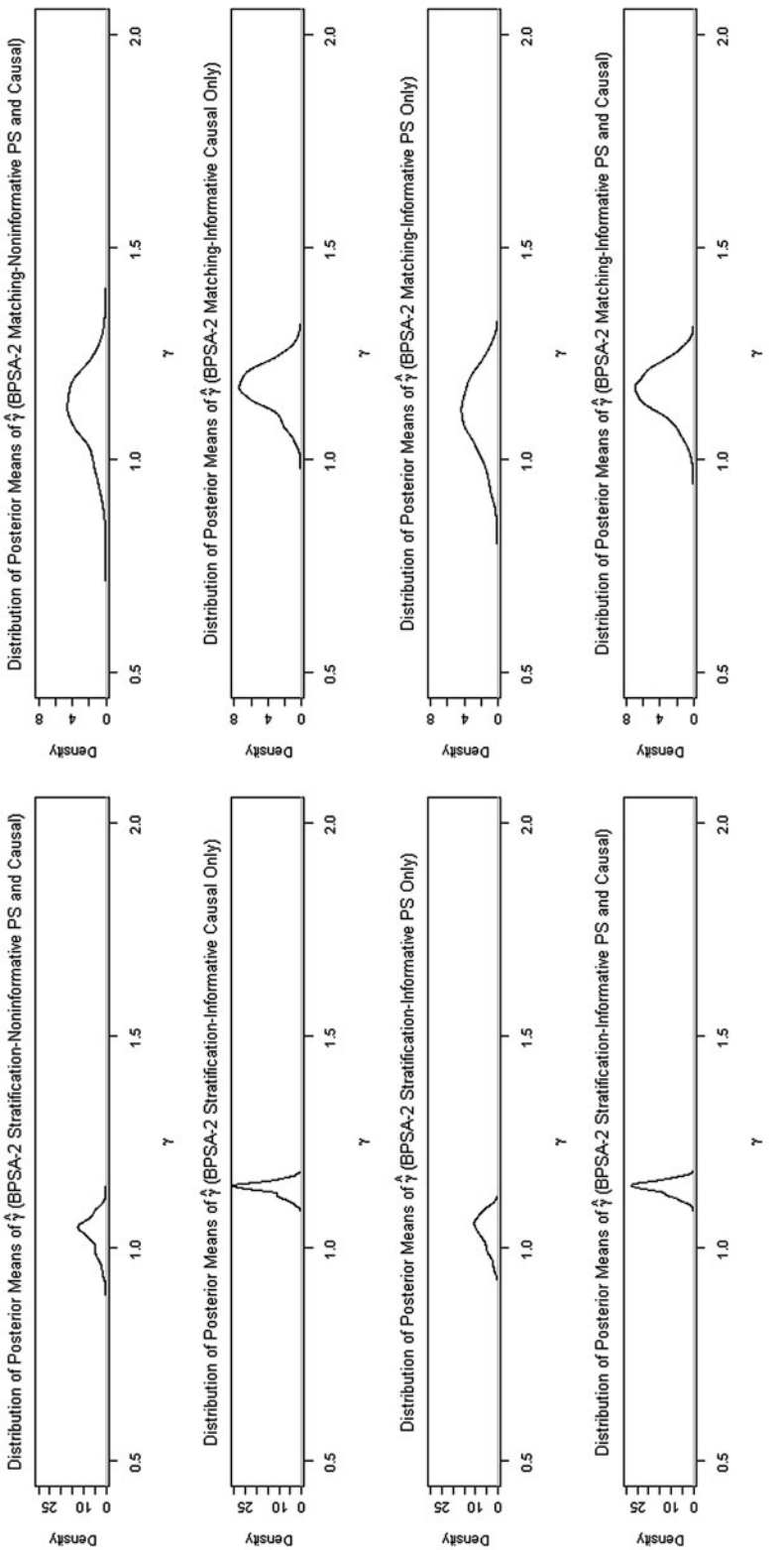

FIGURE 3.
Distribution of posterior means of $\hat{\gamma}$ in BPSA-2 outcome equation.

score model. Compared to stratification, the optimal full matching approach shows a smoother and wider predictive distribution.

To summarize, the results from design I and II show that when applying BPSA, it is desirable to use lower precision when uncertain about the mean in order to obtain estimates similar to frequentist results without adjustment for uncertainty, but with more accurate intervals; or employ higher prior precision when having greater certainty in order to obtain more precise estimates of treatment effects. We also find that the BPSA approach is preferable for small sample sizes, consistent with An's (2010) findings. Among the different methods of propensity score adjustment examined in this paper, the optimal full matching approach often performs better than the stratification method. However, with precise prior information and high precision, the stratification method can also obtain good estimates.

7. A Simulation Study of Standard Error Estimates

To further explore the performance of BPSA-1 and BPSA-2 on the estimated standard error of $\hat{\gamma}$, a small simulation study is conducted, replicating frequentist PSA, BPSA-1 and BPSA-2 with several selected priors. Again, note that when the prior precision of the propensity score equation B_β or the outcome equation B_γ is 0, a non-informative prior for the corresponding equation is used. Whereas when the prior precisions B_β or B_γ are set to 10, the true prior mean is utilized with prior precision 10 for the parameters of the propensity score equation or the outcome equation. The results, averaged over 200 replications, are shown in Table 10. The sample size N is 250 for all approaches shown in Table 10, except $N = 100$ for BPSA-2 optimal full matching (the last four rows).

According to Table 10, the mean of the standard errors for the treatment effect is consistently smaller than the standard deviation of the estimates under stratification, weighting and optimal full matching (0.15 vs. 0.16). The coverage rates for the 95 % confidence intervals are slightly smaller than the nominal level for PSA stratification and weighting (94.5 % and 93.5 %, respectively). Under optimal full matching, the coverage rate is slightly over-estimated (95.5 %).

With informative priors on the propensity score equation, the mean of the estimated standard errors for BPSA-1 achieves the same value as the standard deviation of the estimates under all three propensity score approaches. With non-informative priors, BPSA-1 stratification provides accurate standard errors; however weighting and optimal full matching yield a higher mean standard error compared to the standard deviation of the estimates. The coverage rates for BPSA-1 credible intervals are similar to the frequentist PSA confidence intervals.

For BPSA-2 under stratification and a non-informative prior on the treatment effect, the mean of the estimated standard errors is very precise and the coverage rates are very close to the nominal level. With informative priors on the treatment effect, the mean of the estimated standard errors is higher than the true variability. However, for BPSA-2 under optimal full matching, the mean of the estimated standard errors is much higher than the true variability (0.27 vs. 0.22, 0.21 vs. 0.14) and coverage rates are quite high. This finding may be due to the smaller sample size for BPSA-2 matching. Also, the matching approach seems to yield higher coverage rates under both frequentist and Bayesian methods. In regard to bias, frequentist PSA estimates are as precise or slightly better than BPSA-1 and BPSA-2 estimates. BPSA-2 estimates are closer to true γ when having informative priors on the treatment effect.

In summary, the frequentist PSA variance estimator without taking into account uncertainty in estimated propensity scores tends to underestimate the variance of the treatment effect. On average, BPSA-1 (An's intermediate approach) performs well in terms of the closeness between the estimated and true variability. BPSA-2 stratification provides quite good standard error estimates, but BPSA-2 under optimal full matching over-estimates the variance. Generally, the Bayesian

TABLE 10.
Simulation results for 200 replications.

	Estimate		S.E.		Coverage
	Bias	SD	Mean	SD	
Stratification					
PSA	−0.03	0.16	0.15	0.01	94.5 %
BPSA-1					
$B_\beta = 0$	−0.05	0.16	0.16	0.01	94.5 %
$B_\beta = 10$	−0.05	0.16	0.16	0.01	95.0 %
BPSA-2					
$B_\beta = B_\gamma = 0$	−0.06	0.16	0.16	0.01	94.5 %
$B_\beta = 0$, $B_\gamma = 10$	−0.03	0.07	0.11	0.003	99.0 %
$B_\beta = 10$, $B_\gamma = 0$	−0.06	0.16	0.16	0.01	95.0 %
$B_\beta = B_\gamma = 10$	−0.03	0.07	0.11	0.003	99.0 %
Weighting					
PSA	−0.002	0.16	0.15	0.01	93.5 %
BPSA-1					
$B_\beta = 0$	−0.01	0.16	0.17	0.02	96.5 %
$B_\beta = 10$	−0.04	0.16	0.16	0.01	93.5 %
Optimal matching					
PSA	−0.01	0.16	0.15	0.01	95.5 %
BPSA-1					
$B_\beta = 0$	−0.03	0.16	0.17	0.01	95.5 %
$B_\beta = 10$	−0.03	0.16	0.16	0.01	95.5 %
BPSA-2					
$B_\beta = B_\gamma = 0$	−0.06	0.22	0.27	0.02	97.0 %
$B_\beta = 0$, $B_\gamma = 10$	−0.04	0.14	0.21	0.01	100 %
$B_\beta = 10$, $B_\gamma = 0$	−0.07	0.22	0.27	0.02	97.5 %
$B_\beta = B_\gamma = 10$	−0.04	0.14	0.21	0.01	100 %

Note. Weighting approach in BPSA-2 is not discussed here due to the absence of Bayesian weighted regression in the propensity score literature.

approach has larger coverage rates than the frequentist approach, which may be because the simulation study follows the frequentist framework, thus favoring a frequentist outcome (Yuan & MacKinnon, 2009).

Our simulation results regarding frequentist PSA variance estimation are consistent with McCandless et al. (2009) but disagree with Abadie and Imbens (2009) and An (2010). Assuming group assignment is random, it has been shown that the variance of the estimators using the estimated propensity score is not more than the variance of the estimators given the known true propensity score (Rubin & Thomas, 1992a, 1992b; Abadie & Imbens, 2006). Thus, when treating the estimated propensity score as if it were true, the variance estimator will overestimate the true variance. We believe, therefore, that the discrepancy between our results and those of Abadie and Imbens (2009) and An (2010) could be due to the fact that the asymptotic variance estimator for matching developed by Abadie and Imbens (2006, 2009) treats group assignment as random. In contrast to Abadie and Imbens (2009) and An (2010), the frequentist PSA variance estimator employed by McCandless et al. (2009) and the one developed in this paper treats group assignment as fixed, thereby ignoring the uncertainty of estimated propensity scores, and leading to variance underestimation according to the total variance formula.

From a Bayesian perspective, our two-step Bayesian propensity score approach naturally takes into account the uncertainty of the estimated propensity scores and yields larger variance estimates. Note that our Bayesian results differ from An's (2010) Bayesian results, which may be due to the discrepancy in the methods utilized in the two papers. First, An (2010) focuses on propensity score regression and single nearest neighbor matching methods, whereas our paper studies different approaches, i.e. stratification, weighting and optimal full matching. In addition, An (2010) utilizes an empirical Bayes approach to estimation, whereas we, along with McCandless et al. (2009), implement a more traditional Bayesian approach and fix prior distributions before any data are observed. Finally, the difference between our two-step Bayesian approach and the joint modeling Bayesian approach may also contribute to the discrepancy in results.

8. A Case Study

The real data used for illustrating our method derive from the Early Childhood Longitudinal Study Kindergarten Cohort of 1998 (ECLS-K) (NCES, 2001). The ECLS-K is a nationally representative longitudinal sample providing comprehensive information from children, parents, teachers and schools. The sampled children comes from both public and private schools and attends both full-day and part-day kindergarten programs, having diverse socio-economic and racial/ethnic backgrounds.

In this case study, we examine the treatment effect of full versus part-day kindergarten attendance on IRT-based reading scores for children at the end of 1998 fall kindergarten. A sample of 600 children was randomly selected proportional to the number of children in full- or part-day kindergarten in the population. This resulted in 320 children in full-day kindergarten and 280 children in part-day kindergarten. Thirteen covariates were chosen for the propensity score equation. These included gender, race, child's learning style, self-control, social interactions, sadness/loneliness, impulsiveness/overreactiveness, mother's employment status, whether first time kindergartner in 1998, mother's employment between birth and kindergarten, non-parental care arrangements, social economic status and number of grandparents who live close by. We acknowledge that covariate selection is very important for the use of propensity score method (see e.g. Steiner, Cook, Shadish, & Clark, 2010). However, it is not the focus of this paper and for the purpose of illustration, we assume that the covariates here remain the same before and after treatment assignment and correlate with both the real selection process and the potential outcomes. All analyses utilized the various R software programs described earlier (R Development Core Team, 2011). Missing data were handled via the R program *mice* (multivariate imputation by chained equations) (van Buuren & Groothuis-Oudshoorn, 2011). Non-informative uniform priors are used due to lack of information. The MCMC sampling has 400,000 iterations with burn-in 5000 and thin interval 400, which significantly reduces autocorrelation to an acceptable range.

The estimated treatment effects and confidence/credible intervals for simple regression, traditional PSA, BPSA-1, and BPSA-2 are shown in Table 11. Compared to the nonsignificant results estimated by simple regression, both PSA and BPSA are able to detect a significant treatment effect and greatly reduce the estimation bias. The Bayesian approach with little prior information achieves similar estimated treatment effects to the conventional frequentist approach, but offers a better variance estimate, taking into account the uncertainty of propensity scores and therefore having wider credible intervals. On average, the Bayesian stratification method has 6.2 % wider interval than conventional approach, the Bayesian weighting approach achieves an 8.9 % wider interval, and the Bayesian optimal full matching method obtains as much as 14 %

TABLE 11.
Estimated treatment effects ($\hat{\gamma}$) & standard errors (S.E.) by PSA, BPSA-1 and BPSA-2 for the ECLS-K data.

Method	$\hat{\gamma}$ (S.E.)	Confidence/Credible Interval
Simple linear regression	1.42 (0.78)	(−0.12, 2.95)
Bayesian linear regression	1.43 (0.80)	(−0.14, 2.99)
PSA		
Stratification	2.10 (0.83)	(0.48, 3.72)
Weighting	2.14 (0.80)	(0.57, 3.71)
Optimal matching	2.11 (0.83)	(0.48, 3.74)
BPSA-1		
Stratification	2.08 (0.87)	(0.37, 3.79)
Weighting	2.23 (0.87)	(0.52, 3.94)
Optimal matching	2.08 (0.95)	(0.22, 3.94)
BPSA-2		
Stratification	2.07 (0.88)	(0.34, 3.80)
Optimal matching	2.09 (0.95)	(0.23, 3.94)

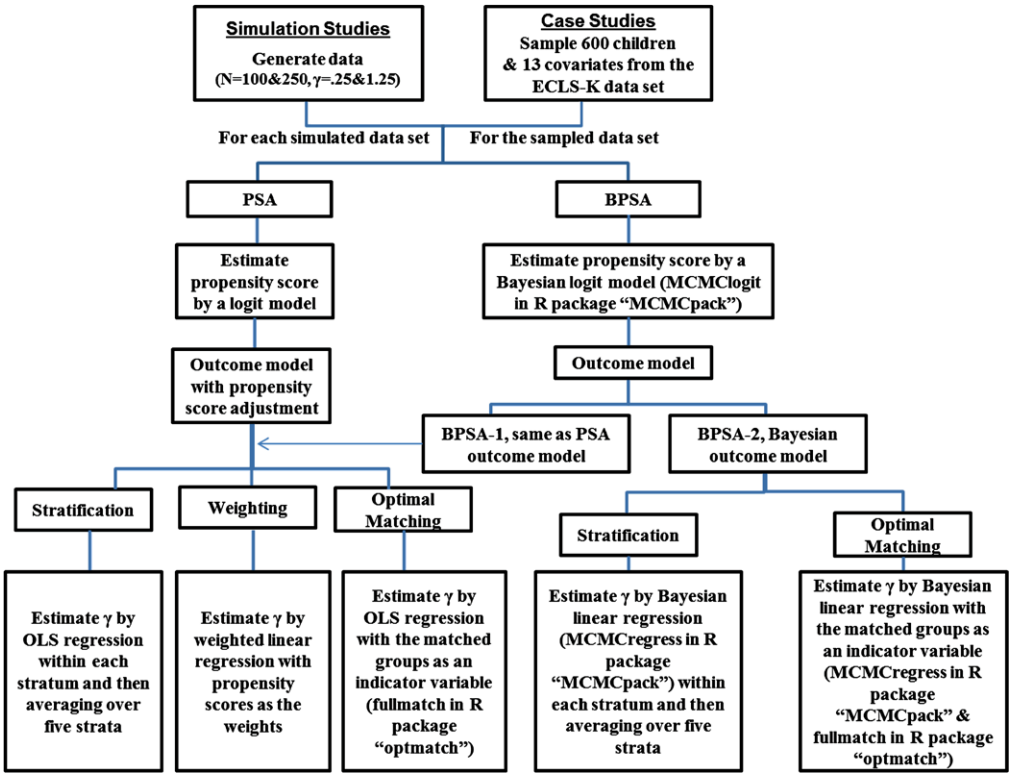

FIGURE 4.
Flowchart outlining steps of simulation and case studies.

wider interval. This result agrees with McCandless et al. (2009) and is consistent with our simulation results and Bayesian theory. A flow chart outlining the steps of the case study can be found on the right hand side of Figure 4.

9. Summary

This paper sought to provide a two-step Bayesian approach to propensity score adjustment that accounts for prior information in both the propensity score model and the outcome model. This approach, here referred to as BPSA-2, was developed for applications to propensity score stratification and optimal full matching. Treatment effect and variance estimators for BPSA-2 under these contexts were also provided. In addition, we extended An's (2010) "intermediate Bayesian approach" (here referred to as BPSA-1) to stratification, weighting, and optimal full matching, as well as provides the treatment effect and variance estimators. Finally, for purposes of completeness we compared our approaches to the conventional frequentist approach which does not account for uncertainty in the propensity scores.

To summarize, simulation study 1 and study 2 examined the general performance of two-step BPSA-1 and BPSA-2 approaches for one simulated data set and explored the influence of various priors on the estimate of a treatment effect. A further simulation study focused on the variance estimation of BPSA-1 and BPSA-2, showing the coverage rates and comparing mean standard errors with approximately true variation. Study 1 revealed that greater precision in the propensity score equation yielded better recovery of the frequentist-based treatment effect compared to traditional PSA and compared to no adjustment. BPSA-1 also revealed a very small advantage to the Bayesian approach for $N = 100$ versus $N = 250$. Design I of the BPSA-2 study revealed that greater precision around the wrong treatment effect can lead to seriously distorted results. Design II of the BPSA-2 study revealed that greater precision around the correct treatment effect yields quite good results, with slight improvement seen with greater precision in the propensity score equation. Our results also indicated that the optimal full matching method obtained somewhat more stable and precise estimates compared to stratification and matching.

An additional simulation study revealed that traditional PSA variance estimates tend to underestimate the true variation of the treatment effect estimators, while BPSA-1 provided accurate or slightly higher variance estimates. For BPSA-2, high prior precision on the treatment effect lead to overly high coverage rates but smaller variation of estimators compared to other approaches. With non-informative priors on the treatment effect, BPSA-2 stratification performed as good as BPSA-1, but BPSA-2 matching provided over-estimated variance estimate and overly high coverage rate that might be due to the smaller sample size. However, this issue deserves further investigation.

We note once again a discrepancy in the findings regarding variance estimation under frequentist PSA. To summarize, Abadie and Imbens (2009) and An (2010) found that the frequentist PSA treatment effect estimator exhibited an overestimation in the variance whereas Lechner (2002), McCandless et al. (2009), and our paper found that the variance estimates of the PSA approach, without taking into account uncertainty of estimated propensity scores, were underestimated. We attribute this discrepancy to the different kinds of PSA variance estimator these papers employ. Abadie and Imbens (2009) and An (2010) consider the PSA variance estimator that has taken into account uncertainty of group assignment, while the latter including this paper utilize the variance estimator that fixes the group assignment. Further research on this issue is warranted.

Finally, the case study revealed that the credible intervals are wider than the confidence intervals when priors are non-informative. This result is consistent with McCandless et al. (2009) and with Bayesian theory.

10. Concluding Remarks

Our paper was situated within the Neyman–Rubin potential outcomes framework, and it is important that we consider the value added of a Bayesian approach to propensity score analysis

within that framework. In particular, in this era of "evidenced-based" research in the social and behavioral sciences and the importance placed on randomized experiments or tightly controlled quasi-experiments, the clear goal is to obtain reliable estimates of treatment effects. Our view is that reliable estimates of treatment effects must account for various types of uncertainty. In the case of a randomized experiment, there is no uncertainty in the mechanism that assigns units to treatment conditions, though there may be considerable uncertainty in the estimate of the treatment effect. This latter uncertainty could be addressed simply by performing a Bayesian regression to test the treatment effect, specifying a lesser or greater degree of precision on the prior distribution of the treatment effect.

In the context of observational studies, however, there are additional sources of uncertainty that need to be addressed. First, there is uncertainty in the choice of covariates to be used in the propensity score equation. The problem of covariate choice has been discussed in Steiner, Cook, and Shadish (2011) and Steiner et al. (2010), who demonstrate important strategies for covariate selection that directly concern the assumption of strong ignorability of treatment assignment. Nevertheless, this source of uncertainty is captured in the disturbance term of the propensity score equation. The second source of uncertainty concerns capturing the correct functional form of the propensity score. Rosenbaum and Rubin (1984, 1985) outline procedures based on iteratively checking covariate balance when including higher-order polynomials and interaction terms in the logistic model for the propensity score.

With respect to our paper, two additional sources of uncertainty can be considered. The third source of uncertainty centers on the degree of prior knowledge we have concerning the propensity score model parameters. This source of uncertainty can be addressed through specification of the prior distributions on the parameters of the propensity score model. Finally, the fourth source of uncertainty centers on the degree of prior knowledge we have on the parameters of the outcome model. This fourth source of uncertainty can be addressed through the specification of the prior distribution on the treatment effect.

Bayesian theory provides the framework by which we can incorporate prior information to represent our degree of uncertainty in the propensity score model and the outcome model. However, careful specification of the prior distribution of all of the model parameters requires deeper understanding of the elicitation problem (see Abbas, Budescu, Yu, & Haggerty, 2008; Abbas, Budescu, & Gu, 2010; O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, & et al., 2006). The general idea is that through a careful review of prior research on a problem, and/or the careful elicitation of prior knowledge from content area experts and/or key stakeholders, relatively precise values for hyperparameters can be obtained and incorporated into a Bayesian specification. Alternative elicitations can be directly compared via Bayesian model selection measures such as the deviance information criterion (Spiegelhalter et al., 2002). It is through (a) the careful and rigorous elicitation of prior knowledge, (b) the incorporation of that knowledge into the propensity score model and outcome model, and (c) a rigorous approach to the selection among competing propensity score models, that a pragmatic *and* evolutionary development of knowledge can be realized—and this is precisely the advantage that Bayesian statistics, and Bayesian propensity score analysis in particular, has over its frequentist counterparts. Now that the theoretical and computational foundations have been established, the benefits of Bayesian propensity score analysis in the context of observational studies will be realized in terms of how it provides insights into important substantive problems.

Acknowledgements

The research reported in this paper was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D110001 to The University of

Wisconsin–Madison. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Abadie, A., & Imbens, G. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*, 235–267.

Abadie, A., & Imbens, G.W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, *76*, 1537–1558.

Abadie, A., & Imbens, G.W. (2009). *Matching on the estimated propensity score* (NBER Working Paper 15301).

Abbas, A.E., Budescu, D.V., & Gu, Y. (2010). Assessing joint distributions with isoprobability countours. *Management Science*, *56*, 997–1011.

Abbas, A.E., Budescu, D.V., Yu, H.T., & Haggerty, R. (2008). A comparison of two probability encoding methods: fixed probability vs. fixed variable values. *Decision Analysis*, *5*, 190–202.

An, W. (2010). Bayesian propensity score estimators: incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, *40*, 151–189.

Austin, P.C., & Mamdani, M.M. (2006). A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine*, *25*, 2084–2106.

Benjamin, D.J. (2003). Does 401(k) eligibility increase saving? Evidence from propensity score subclassification. *Journal of Public Economics*, *87*, 1259–1290.

Cochran, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*, 295–313.

Dawid, A.P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, *77*, 605–610.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2003). *Bayesian data analysis* (2nd ed.). London: Chapman and Hall.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Guo, S., & Fraser, M.W. (2010). *Propensity score analysis: statistical methods and applications*. Thousand Oaks: Sage.

Hansen, B.B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, *99*, 609–618.

Hansen, B.B., & Klopfer, S.O. (2006). Optimal full matching and related designs via network flow. *Journal of Computational and Graphical Statistics*, *15*, 609–627.

Heckman, J.J. (2005). The scientific model of causality. In R.M. Stolzenberg (Ed.), *Sociological methodology* (Vol. 35, pp. 1–97). Boston: Blackwell Publishing.

Hirano, K., & Imbens, G.W. (2001). Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, *2*, 259–278.

Hirano, K., Imbens, G.W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*, 1169–1189.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Horvitz, D.G., & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*, 663–685.

Hoshino, T. (2008). A Bayesian propensity score adjustment for latent variable modeling and MCMC algorithm. *Computational Statistics & Data Analysis*, *52*, 1413–1429.

Larsen, M.D. (1999). An analysis of survey data on smoking using propensity scores. *Sankya. The Indian Journal of Statistics*, *61*, 91–105.

Lechner, M. (2002). Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society. Series A. Statistics in Society*, *165*, 59–82.

Lunceford, J.K., & 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.

Martin, A.D., Quinn, K.M., & Park, J.H. (2010, May 10). Markov chain Monte Carlo (MCMC) package. http://mcmcpack.wustl.edu/.

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

NCES (2001). *Early childhood longitudinal study: kindergarten class of 1998–99: base year public-use data files user's manual* (Tech. Rep. No. NCES 2001-029). U.S. Government Printing Office.

Neyman, J.S. (1923). Statistical problems in agriculture experiments. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *2*, 107–180.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., et al. (2006). *Uncertain judgements: eliciting experts' probabilities*. West Sussex: Wiley.

Perkins, S.M., Tu, W., Underhill, M.G., Zhou, X.H., & Murray, M.D. (2000). The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety*, *9*, 93–101.

R Development Core Team (2011). *R: a language and environment for statistical computing* (Computer software manual). Vienna, Austria. Available from http://www.R-project.org (ISBN 3-900051-07-0).

Rässler, S. (2002). *Statistical matching: a frequentist theory, practical applications, and alternative Bayesian approaches*. New York: Springer.

Rosenbaum, P.R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*, 387–394.
Rosenbaum, P.R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*, 1024–1032.
Rosenbaum, P.R. (2002). *Observational studies* (2nd ed.). New York: Springer.
Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.
Rosenbaum, P.R., & Rubin, D.B. (1984). Reducing bias in observational studies using sub-classification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.
Rosenbaum, P.R., & Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate a propensity score. *American Statistician*, *39*, 33–38.
Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.
Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference. *Bayesian Statistics*, *2*, 463–472.
Rubin, D.B. (2006). *Matched sampling for causal effects*. Cambridge: Cambridge University Press.
Rubin, D.B., & Thomas, N. (1992a). Affinely invariant matching methods with ellipsoidal distributions. *Annals of Statistics*, *20*, 1079–1093.
Rubin, D.B., & Thomas, N. (1992b). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*, *79*, 797–809.
Rubin, D.B., & Thomas, N. (1996). Matching using estimated propensity scores. *Biometrics*, *52*, 249–264.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *64*, 583–639.
Steiner, P.M., & Cook, D. (in press). Matching and propensity scores. In T. Little (Ed.), Oxford handbook of quantitative methods. Oxford: Oxford University Press.
Steiner, P.M., Cook, T.D., & Shadish, W.R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, *36*, 213–236.
Steiner, P.M., Cook, T.D., Shadish, W.R., & Clark, M.H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*, 250–267.
Thoemmes, F.J., & Kim, E.S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*, 90–118.
van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. Available from http://www.jstatsoft.org/v45/i03/.
Yuan, Y., & MacKinnon, D.P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*, 301–322.
Zanutto, E.L., Lu, B., & Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national anti-drug media campaign. *Journal of Educational and Behavioral Statistics*, *30*, 59–73.

Manuscript Received: 14 OCT 2010
Final Version Received: 5 OCT 2011