

This article was downloaded by: [McGill University Library]

On: 29 May 2015, At: 10:35

Publisher: Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Journal of the American Statistical Association

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/uasa20>

Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model-Averaged Causal Effects

Corwin Matthew Zigler & Francesca Dominici

Accepted author version posted online: 18 Dec 2013. Published online: 19 Mar 2014.



CrossMark

[Click for updates](#)

To cite this article: Corwin Matthew Zigler & Francesca Dominici (2014) Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model-Averaged Causal Effects, Journal of the American Statistical Association, 109:505, 95-107, DOI: [10.1080/01621459.2013.869498](https://doi.org/10.1080/01621459.2013.869498)

To link to this article: <http://dx.doi.org/10.1080/01621459.2013.869498>

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at <http://www.tandfonline.com/page/terms-and-conditions>

Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model-Averaged Causal Effects

Corwin Matthew ZIGLER and Francesca DOMINICI

Causal inference with observational data frequently relies on the notion of the propensity score (PS) to adjust treatment comparisons for observed confounding factors. As decisions in the era of “big data” are increasingly reliant on large and complex collections of digital data, researchers are frequently confronted with decisions regarding which of a high-dimensional covariate set to include in the PS model to satisfy the assumptions necessary for estimating average causal effects. Typically, simple or ad hoc methods are employed to arrive at a single PS model, without acknowledging the uncertainty associated with the model selection. We propose three Bayesian methods for PS variable selection and model averaging that (a) select relevant variables from a set of candidate variables to include in the PS model and (b) estimate causal treatment effects as weighted averages of estimates under different PS models. The associated weight for each PS model reflects the data-driven support for that model’s ability to adjust for the necessary variables. We illustrate features of our proposed approaches with a simulation study, and ultimately use our methods to compare the effectiveness of surgical versus nonsurgical treatment for brain tumors among 2606 Medicare beneficiaries. Supplementary materials for this article are available online.

KEY WORDS: Bayesian statistics; Causal inference; Comparative effectiveness; Model averaging.

1. INTRODUCTION

Causal inference with observational data frequently relies on methods based on the propensity score (PS; Rosenbaum and Rubin 1983), as these methods are designed to estimate causal treatment effects by adjusting for observed confounding factors. PS methods typically unfold in two stages: first, a model is fit to estimate the probability of assignment to treatment (the estimated PS), and then outcomes of interest are compared between treated and untreated units having similar values of the estimated PS. As with all methods for making causal inferences with observational data, PS methods require that the researchers specify the confounders necessary to satisfy the assumption of strong ignorability (i.e., the “no unmeasured confounding” assumption), rendering correct specification of the PS model of vital importance. However, the correct set of variables to include in the PS model is rarely known, and researchers in the modern era of “big data” are increasingly confronted with decisions regarding which of many covariates to include in the PS. Such is often the case when, for example, large administrative databases are used to compare the effectiveness of medical therapies as they are employed in routine clinical practice.

One recommendation for choosing a PS model is to simply include every available covariate. Aside from the efficiency sacrificed by the inclusion of possibly extraneous variables, this “kitchen sink” approach is becoming increasingly untenable as PS methods are deployed in high-dimensional data settings. In attempt to include only a subset of relevant variables in the PS model, researchers often combine expert knowledge with tradi-

tional or ad hoc procedures to select variables that are predictive of treatment assignment (Austin 2008). Methods targeting the subset of variables that best predicts treatment assignment have noted drawbacks in terms of efficiency and finite-sample bias. In fact, the target set of variables to include in the PS model should also take into account how covariates are related to the outcome (Rubin and Thomas 1996; Rubin 1997; Brookhart et al. 2006). Despite disagreement about whether data-driven approaches to select the PS model should incorporate information in the outcome (Rubin 2008), the purported benefit of including important outcome predictors has spawned data-driven methods to do so, such as the reasoned algorithmic approach of Schneeweiss et al. (2009).

Regardless of the exact procedure used, current methods for selecting PS models fail to account for the uncertainty in the selection, and have never been considered within a Bayesian framework. Once a researcher arrives at a particular PS model, all subsequent inference conditions on this single model. The lack of methods to formally acknowledge the uncertainty in PS model selection is an important barrier to tailoring PS methods to modern research priorities such as the Big Data Research and Development Initiative put forth by the United States government. To acknowledge the inherent uncertainty in the specification of the PS model, we propose three Bayesian procedures for PS models that (a) select relevant variables from a (potentially large) set of candidate variables to include in the PS and (b) estimate causal treatment effects as weighted averages of estimates under different PS models. One vital feature of our approach is that the associated weight for each PS model reflects the data-driven support for that model’s ability to adjust for the desired variables. Thus, our methods account for the uncertainty associated with what is arguably the most important choice in a PS analysis: the choice of which variables to include in the PS.

Corwin M. Zigler is Assistant Professor (E-mail: czigler@hsph.harvard.edu) and Francesca Dominici is Professor of Biostatistics (E-mail: fdominic@hsph.harvard.edu), Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115. This work is supported by NCI P01 CA134294, EPA RD 83479081, R834893, HEI 4909-RFA11-1/12-3. The contents of this work are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication. The authors thank Giovanni Parmigiani, Chi Wang, Sebastien Haneuse, Matt Cefalu, and Tom Belin for helpful comments and discussion.

Our proposed methodology has important points of contact with the recently proposed method for “Bayesian Adjustment for Confounding” (BAC), which Wang, Parmigiani, and Dominici (2012) introduced as a means for selecting variables to include in a regression model based on associations with both an exposure and an outcome. Although Wang, Parmigiani, and Dominici (2012) did not explicitly consider the PS, discussions of that work pointed toward similar notions of Bayesian confounder selection in PS models (McCandless 2012; Vansteelandt 2012), which we explore in detail here. In founding our approach from a Bayesian perspective, our methodology is a continuation of recent work investigating Bayesian PS estimation based on a single joint likelihood representing both the PS and outcome models simultaneously (McCandless, Gustafson, and Austin 2009; Zigler et al. 2013), which is distinct from traditional sequential PS approaches that first estimate the PS model to obtain the estimated PS, then treat these estimated quantities as fixed and known for estimation of the outcome model.

Section 2 of this article reviews PS methods and draws distinctions between traditional sequential and joint Bayesian estimation. Section 3 outlines three related approaches for Bayesian variable selection and model averaging for PS methods. Section 4 presents a simulation study to illustrate the key features of our proposed approaches. Section 5 compares the effectiveness of surgical versus nonsurgical treatment for malignant brain tumors among 2606 Medicare beneficiaries. We conclude with a discussion.

2. BAYESIAN PROPENSITY SCORE ESTIMATION

In this section, we briefly review the motivation for PS methods and summarize joint Bayesian PS estimation as distinct from traditional sequential procedures. We consider likelihood-based methods that estimate causal effects with a model for the outcome that adjusts for the PS.

2.1 Estimating Causal Effects With Propensity Scores

For the i th observational unit, denote the binary treatment with $X_i = 0, 1$, the outcome of interest with Y_i , and a vector of p measured pretreatment covariates with $U_i = (U_{i1}, U_{i2}, \dots, U_{ip})$. Rosenbaum and Rubin (1983) defined the PS as the conditional probability of assignment to treatment $X = 1$, given U , and illustrated that the PS enjoys the properties of a *balancing score* in that, conditional on the PS, the individual covariates are orthogonal to the treatment indicator: $X \perp U | \text{PS}$. The balancing property of the PS combined with the assumption of strongly ignorable treatment assignment (i.e., that there are no unmeasured confounders) implies that outcome comparisons between treated and untreated units with the same value of the PS represent estimates of average causal effects (ACEs; Rosenbaum and Rubin 1983). One salient benefit of estimating causal effects in this manner is that it does not necessarily rely on a detailed parametric model for the outcome. Rather, covariate information is condensed into a scalar quantity (the PS), adjustment for which can recover causal estimates without strong parametric assumptions pertaining to how covariates relate to Y .

2.2 Models for Propensity Score Estimation and the Average Causal Effect

Typically, the PS, defined as $E[X|U]$, is estimated with a model for the relationship between X and U , which we represent with the generalized linear model:

$$g_x(E[X_i|U_i]) = \sum_{k=0}^p \alpha_k \gamma_k U_{ik}, \quad (1)$$

where we define $U_{i0} \equiv 1$ for all i to denote a model intercept. Here, γ_k ($k \geq 1$) represents the coefficient describing the association between U_k and the probability of treatment assignment to $X = 1$, and $g_x(\cdot)$ is a link function. An individual PS model is denoted with a particular value of the vector α , denoted $\alpha^m = (\alpha_0^m, \alpha_1^m, \alpha_2^m, \dots, \alpha_p^m)$, where each $\alpha_k^m = 1, 0$ represents whether U_k is or is not included in the m th PS model. Throughout, we fix $\alpha_0 = 1$ to force the inclusion of a model intercept. For brevity, we refer to a particular α^m as a particular model, and say that U_k with $\alpha_k^m = 1$ are variables included in model α^m . In general, a particular U_k could represent a derived function such as an interaction or higher order polynomial term.

Estimating causal effects with the PS often relies on a model for the outcome, conditional on the PS. While models for general outcome types are permitted, we simplify notation by considering only generalized linear models for binary outcomes of the form

$$g_y(E[Y_i|X_i, U]) = \beta_0 + \beta_X X_i + h(\text{PS}(\gamma, \alpha, U_i); \xi) + \sum_{k=1}^p \alpha_k \delta_k U_{ik}, \quad (2)$$

where $Y_i = 0, 1$, $g_y(\cdot)$ is a link function, and the notation $\text{PS}(\gamma, \alpha, U_i)$ explicates that the PS for the i th unit is a deterministic function of (γ, α, U_i) , that is, $\text{PS}(\gamma, \alpha, U_i) = g_x^{-1}(\sum \alpha_k \gamma_k U_{ik})$. For this outcome model, β_X represents the conditional treatment effect at a given value of the PS, and treatment-by-PS interactions could be included. The function $h(\text{PS}(\gamma, \alpha, U_i); \xi)$ denotes how the PS enters the outcome model depending on unknown parameters ξ , and the $\sum \alpha_k \delta_k U_{ik}$ represents residual adjustment for individual covariates in addition to the PS. Note that the $\text{PS}(\gamma, \alpha, U)$ and the residual adjustment term $\sum \alpha_k \delta_k U_{ik}$ depend on the same α , implying that, for a given α^m , the PS is comprised of the variables in α^m , and residual adjustment for an individual U_k is conducted in the outcome model if and only if that variable is included in α^m . For example, $h(\text{PS}(\gamma, \alpha, U_i); \xi)$ could denote dummy variables indicating membership in subclasses defined by quantiles of the PS, with $\sum \alpha_k \delta_k U_{ik}$ denoting linear adjustment for each covariate included in the PS. We detail the rationale for including the residual adjustment $\sum \alpha_k \delta_k U_{ik}$ in subsequent sections, but note here that this strategy is akin to doubly robust procedures that will estimate causal effects if either the PS model or the residual adjustment is correctly specified (Little and An 2004; Bang and Robins 2005).

Upon specification of the U required to satisfy the assumption of strongly ignorable treatment assignment, the ACE is defined as $\Delta = E_U\{E[Y|X = 1, U] - E[Y|X = 0, U]\}$, which, provided that α contains the necessary U , can be obtained by calculating the average differences between the predicted

values from model (2) with X set to 1 and the analogous predicted values with X set to 0. We revisit the implications of estimating the ACE under different values of α in Section 3.

2.3 Joint Bayesian PS Estimation

Traditional PS estimation is carried out sequentially in the sense that the researcher first specifies the variables to include in the PS (i.e., sets $\alpha = \alpha^0$) and estimates γ from (1). Then, the selected α^0 , the estimated $\hat{\gamma}$, and implied estimated PS are used for estimation of (2), that is, estimation of causal effects is based on $g_y(E[Y_i|X_i, U_i]) = \beta_0 + \beta_X X_i + h(\text{PS}(\hat{\gamma}, \alpha^0, U_i); \xi) + \sum \alpha_k^0 \delta_k U_{ik}$. Several limitations of this sequential approach have been noted, including the misstatement of uncertainty due to treating $\hat{\gamma}$ (and hence, the estimated PS) as fixed and known in the outcome model, when these quantities are in fact estimated with error (Gelman and Hill 2007). More pertinent to model uncertainty, the decision regarding which variables to include in the PS (i.e., setting $\alpha = \alpha^0$) is made in the first stage and also treated as fixed in the second stage.

In contrast to sequential methods, joint Bayesian methods have been recently introduced to estimate causal effects with a prespecified $\alpha = \alpha^0$ (McCandless, Gustafson, and Austin 2009; Zigler et al. 2013). Extending these methods to settings with unknown α , joint Bayesian inference relies on the following likelihood that simultaneously considers quantities in the PS and outcome models:

$$L(\mathbf{Y}, \mathbf{X}|\mathbf{U}, \gamma, \alpha, \beta, \delta) = \prod_{i=1}^n \left\{ \left[g_x^{-1} \left(\sum_{k=0}^p \alpha_k \gamma_k U_{ik} \right) \right]^{X_i} \times \left[1 - g_x^{-1} \left(\sum_{k=0}^p \alpha_k \gamma_k U_{ik} \right) \right]^{1-X_i} \right. \\ \times \left[g_y^{-1} (\beta_0 + \beta_X X_i + h(\text{PS}(\gamma, \alpha, U_i); \xi) + \sum_{k=1}^p \alpha_k \delta_k U_{ik}) \right]^{Y_i} \\ \times \left[1 - g_y^{-1} (\beta_0 + \beta_X X_i + h(\text{PS}(\gamma, \alpha, U_i); \xi) + \sum_{k=1}^p \alpha_k \delta_k U_{ik}) \right]^{1-Y_i} \left. \right\}, \quad (4)$$

where, here and throughout, boldface represents vectors and matrices for the entire sample of n units. This likelihood, coupled with a prior distribution $p(\gamma, \alpha, \beta, \xi, \delta)$ serves as the basis for joint Bayesian estimation of causal effects. One key feature of joint Bayesian estimation with the likelihood in (3) and (4) is that the α and γ determining the PS are treated as unknown quantities, uncertainty about which is integrated out of posterior distributions of causal effects. Not only does this propagate uncertainty regarding the variables contained in the PS model and the estimates of the PS themselves, but it also allows information from the outcome model to contribute to the PS model.

2.4 Model Feedback

Note that some parameters (namely, γ and α) appear in both term (3) and term (4) of the above likelihood. As a result, posterior distributions of these parameters depend on both the PS and outcome models, a phenomenon referred to as “feedback”

because quantities in the outcome model will indirectly inform quantities in the PS model. With the likelihood in (3) and (4), there are two distinct sources of feedback. The first relates to the appearance of γ in both terms of the likelihood, meaning that Y and other quantities in the outcome model inform estimation of γ and hence, the PS. This type of feedback, which we refer to here as γ -feedback, has been previously considered when α is fixed and known, and was shown to be potentially detrimental to PS estimation (Zigler et al. 2013). Specifically, Zigler et al. (2013) showed that, conditional on any model α , γ -feedback from the outcome model into the PS model can distort the balancing-score property of the PS and yield biased estimates of causal effects. Zigler et al. (2013) also showed that one way to prevent distortion of the PS is to conduct residual adjustment for every covariate that appears in the PS model. In other words, when conducting joint Bayesian PS estimation, we specify the outcome model (2) to include residual adjustment via $\sum \alpha_k \delta_k U_{ik}$.

The inclusion of the parameter α in terms (3) and (4) of the likelihood leads to another source of feedback. We refer to this feedback as α -feedback which, in principle, will allow variable selection to be conducted based jointly on covariate associations with X and with Y . The use of α -feedback to select variables based on associations with both X and Y is an important departure from traditional PS approaches. Note here that α feedback would occur even without residual adjustment with $\sum \alpha_k \delta_k U_{ik}$ because α implicitly appears in the outcome model via $h(\text{PS}(\gamma, \alpha, U_i); \xi)$.

Either source of feedback could also be “cut” by conducting an approximately Bayesian analysis that still uses the likelihood in (3) and (4), but updates some parameters from conditional posterior distributions that exclude some terms of the likelihood and prior distributions (Liu, Bayarri, and Berger 2009; Lunn et al. 2009). For example, updating α from $p(\alpha|X, U, \gamma)$ would “cut the feedback” between the outcome model and the selection of variables. We detail one such approximately Bayesian approach in Section 3.2.

3. BAYESIAN VARIABLE SELECTION AND MODEL AVERAGING FOR PROPENSITY SCORES

As alluded to in Section 2.2, the introduction of the parameter α denoting whether each variable is included in the PS model reflects the model uncertainty. Essentially, by adding the parameter α , we can estimate the causal treatment effect as a weighted average over different PS and outcome models, with weights corresponding to data-driven support of whether each variable should or should not be included for adjustment (Hoeting et al. 1999).

More formally, let \mathcal{M} denote the set of all models being considered, which, for our purposes, will consist of the $M = 2^p$ possible values of $\alpha^m = (1, \alpha_1^m, \alpha_2^m, \dots, \alpha_p^m)$ (recall α_0 is fixed to 1). With prior probability of model m denoted as $p(\alpha^m)$, the posterior probability of model m is

$$p(\alpha^m|\text{Data}) = \frac{p(\alpha^m)p(\text{Data}|\alpha^m)}{\sum_{\alpha^i \in \mathcal{M}} p(\alpha^i)p(\text{Data}|\alpha^i)}.$$

Our goal is estimation of the ACE of treatment with $X = 1$ versus treatment with $X = 0$, making our target for inference

the posterior distribution of the ACE, which will be a weighted average of estimates of Δ under each model in \mathcal{M} , with weights corresponding to the posterior probability of each model:

$$p(\Delta|\text{Data}) \approx \sum_{\alpha^m \in \mathcal{M}} p(\Delta^{\alpha^m}|\alpha^m, \text{Data})p(\alpha^m|\text{Data}), \quad (5)$$

where $\Delta^{\alpha^m} = E_{U_{\alpha^m}}\{E[Y|X = 1, U_{\alpha^m}] - E[Y|X = 0, U_{\alpha^m}]\}$ and U_{α^m} denotes the subset of U in α^m . Note that Δ^{α^m} will not necessarily share the same interpretation as the causal effect Δ , resulting in the approximation in (5). The key issue for interpreting Δ^{α^m} as an estimate of the causal effect pertains to whether α^m contains the confounders necessary to satisfy the assumption of strong ignorability; failure to include even one confounder would result in a Δ^{α^m} that does not share the causal interpretation of Δ . In theory, there exists some minimal model, α^* , that is sufficient for satisfying the assumption of strong ignorability and interpreting Δ^{α^*} as the causal estimate. Adding variables to α^* will not alter this causal interpretation. We treat α^* as unknown, assuming only that it is comprised of a subset of the p variables available for analysis. The motivation for using (5) for estimating causal effects is that the procedures outlined below assign posterior weights, $p(\alpha^m|\text{Data})$, based jointly on covariate associations with X and Y or on covariate associations with X only. This approximates the true posterior distribution of Δ by distributing most posterior weight to models containing α^* . Similar reasoning motivates the BAC method of Wang, Parmigiani, and Dominici (2012).

In practice, evaluation of each model's posterior probability may be infeasible for all M models under consideration (e.g., $p = 20$ implies over 1 million models). In such cases, Markov chain Monte Carlo (MCMC) algorithms that sample from the joint posterior of $p(\alpha, \Delta|\text{Data})$ have been developed to estimate quantities such as that in (5) by visiting only the models with nonnegligible posterior support from the data (George and McCulloch 1997; Hoeting et al. 1999). We explore two such methods in this article: the MCMC Model Composition (MC³) approach of Madigan, York, and Allard (1995) and the stochastic search variable selection (SSVS) method of George and McCulloch (1997). Throughout, we assume equal prior probability for all M models.

3.1 Bayesian PS Model-Averaged Causal Effects With MC³

Given a current value for α^m , the MC³ method constructs an MCMC chain that iteratively visits “neighboring” models by either adding or removing a single variable from α^m to obtain α' , then moving the chain by setting $\alpha = \alpha'$ with probability $\min\{1, \frac{p(\alpha'|\text{Data})}{p(\alpha^m|\text{Data})}\}$ and estimating Δ^α (Madigan, York, and Allard 1995; Raftery, Madigan, and Hoeting 1997). One key necessity for this approach is an available expression for $p(\alpha^m|\text{Data})$, which requires an analytically tractable expression for the integrated likelihood: $p(\text{Data}|\alpha^m) = \int p(\text{Data}|\alpha^m, \theta)p(\theta|\alpha^m)d\theta$, where θ here represents the collection of all parameters for which presence or absence in the model depends on α (e.g., $\theta = (\gamma, \delta)$ in (3) and (4)). While the integrated likelihood is readily available for some standard problems (e.g., linear regression), it will only be available in closed form for the PS

likelihood in (3) and (4) under specific parametric assumptions, which is illustrated in Appendix A.2. In particular, for fully Bayesian estimation with MC³, we assume

- (i) $g_x(\cdot)$ and $g_y(\cdot)$ are both Φ^{-1} representing Probit regression in both the PS and the outcome models.
- (ii) $h(\text{PS}(\gamma, \alpha, U_i); \xi) = \xi \times \sum \alpha_k \gamma_k U_{ik}$, that is, the outcome model in (2) adjusts for $\Phi^{-1}(\text{PS})$ as a linear covariate.
- (iii) Residual adjustment via $\sum \alpha_k \delta_k U_{ik}$ is included with one variable having $\alpha_k = 1$ removed to prevent perfect collinearity with the PS.

Incorporating these assumptions into the likelihood as expressed in (3) and (4) essentially negates one integral feature using the PS to adjust for confounding. Specifically, assumptions (ii) and (iii) yield an outcome model that is a simple reparameterization of a model containing a linear main effect term for every variable in α but *not the propensity score*. This is shown in Appendix A, along with details of the prior specification and complete MCMC algorithm.

Conducting the above MC³ procedure is operationally a PS approach in that it relies on specification of a PS model and outcome model with (3) and (4), but the restrictive parametric model implied by assumptions (i)–(iii) is a Probit regression with only main effect terms for each covariate; precisely the type of model that PS methods are typically used to avoid. Nonetheless, this strategy will be useful for illustrating (via simulation studies in Section 4) the salient issues involved in selecting variables for PS models.

3.2 Approximately Bayesian PS Model-Averaged Causal Effects With MC³

Altering the MC³ approach of Section 3.1 to allow more flexible model specifications can be accomplished with an approximately Bayesian analysis of the likelihood in (3) and (4) by “cutting the feedback” (or modularizing) the PS and outcome models (Liu, Bayarri, and Berger 2009; Lunn et al. 2009). Specifically, we propose an approximately Bayesian MC³ approach, where, at each MCMC iteration, the ratio $\frac{p(\alpha'|\text{Data})}{p(\alpha^m|\text{Data})}$ governing acceptance of a move between models is obtained without regard to Y , as are updates of γ and the PS. Then, (β, ξ, δ) are simulated conditional on α and the PS, which only involves term (4) of the likelihood and the relevant prior distributions. Full details of the approximately Bayesian method appear in Appendix A.4. The key benefit of this approach is that the integrated likelihood used to calculate $\frac{p(\alpha'|\text{Data})}{p(\alpha^m|\text{Data})}$ only involves integration over quantities in the PS model, and, for Probit regression, this integration can be easily evaluated without assumptions (ii) and (iii) that were required for the fully Bayesian MC³ approach in Section 3.1. The absence of γ -feedback in the approximately Bayesian method also allows omission of residual adjustment term $\sum \alpha_k \delta_k U_{ik}$ from (2), if desired. However, the flexibility gained by cutting feedback between the PS and outcome model comes at the cost of foregoing the use of Y to inform the selection of α , meaning that inclusion of variables in the PS model is only informed by associations with X .

3.3 Bayesian PS Model-Averaged Causal Effects With SSVS

In contrast to the fully Bayesian method of Section 3.1 that entails restrictive parametric assumptions and the approximately Bayesian approach in Section 3.2 that precludes Y from informing the PS model, we also propose a fully Bayesian SSVS approach with so-called “spike-and-slab” hierarchical mixture priors for the coefficients in the PS and outcome models (George and McCulloch 1997). This strategy reformulates the expressions in (1) and (2) to yield the following:

$$g_x(E[X_i|U_i]) = \sum_{k=0}^p \gamma_k U_{ik}, \quad (6)$$

$$g_y(E[Y_i|X_i, U]) = \beta_0 + \beta_X X_i + h(\text{PS}(\gamma, U_i); \xi) + \sum_{k=1}^p \delta_k U_{ik}, \quad (7)$$

$$p(\gamma_k|\alpha_k) = N(0, \tau_\gamma^2) \times \alpha_k + N(0, c_\gamma^{-2}\tau_\gamma^2) \times (1 - \alpha_k), \quad (8)$$

$$p(\delta_k|\alpha_k) = N(0, \tau_\delta^2) \times \alpha_k + N(0, c_\delta^{-2}\tau_\delta^2) \times (1 - \alpha_k), \quad (9)$$

for $k = 0, 1, \dots, p$, where $\tau_\gamma^2, \tau_\delta^2$ are hyperparameters chosen to reflect proper but noninformative prior distributions for (γ_k, δ_k) with $\alpha_k = 1$ (the “slab”), and (c_γ, c_δ) are hyperparameters chosen to concentrate prior mass very tightly around zero for (γ_k, δ_k) with $\alpha_k = 0$ (the “spike”). Rather than completely remove variables from the model when $\alpha_k = 0$, the SSVS approach shrinks coefficients very close to zero. Note that the PS setting involves the unique feature that the mixture priors are specified for two sets of parameters (γ and δ) but, for a given k , both γ_k and δ_k are constrained to belong to the same mixture component because there is only one model indicator, namely, α .

As compared to the MC³ approaches in Sections 3.1 and 3.2, the SSVS approach does not require any additional parametric assumption. Model selection proceeds by iteratively sampling $\gamma_k, \delta_k, \beta$, and ξ from their appropriate conditional posterior distributions, then updating each α_k from a binomial distribution that depends on the current values of γ_k and δ_k and on the prior distributions in (8) and (9). Note that while the inclusion of the residual adjustment term $\sum \alpha_k \delta_k U_{ik}$ is primarily motivated by the discussion of γ -feedback in Section 2.4, this residual adjustment is required for α feedback in the SSVS approach so that $p(\alpha_k = 1|\text{Data})$ will be closer to one for variables with $\gamma_k^2 + \delta_k^2$ large. Thus, unlike the MC³ approaches in Sections 3.1 and 3.2, the SSVS approach incorporates outcome information into the PS model selection while also allowing flexible parametric specification. These benefits come at the cost of some practical difficulties in implementation. Specifically, the choice of hyperparameters for the mixture priors in (8) and (9) will affect which variables are selected to have $\alpha_k = 1$, and SSVS with the likelihood in (3) and (4) can be computationally expensive for large values of p . Full details appear in the Web Appendix A. To simplify our presentation, we henceforth fix $\tau_\gamma = \tau_\delta = \tau$ and $c_\gamma = c_\delta = c$.

4. SIMULATIONS TO ILLUSTRATE PROPENSITY SCORE VARIABLE SELECTION AND MODEL AVERAGING

Here, we use simulated data to illustrate the approaches for Bayesian variable selection and model averaging in Section 3. For all illustrations, data are generated as follows: for $i = 1, 2, \dots, n$, we simulate p covariates $U_i = (U_{i1}, U_{i2}, \dots, U_{ip})$ from $\text{MVN}(0, I)$, where I is the identity matrix, and U_{0i} is set to 1 to denote a model intercept. X_i is simulated from a Bernoulli distribution with probability $X_i = 1$ specified with (1), with $g_x(\cdot)$ set to $\Phi^{-1}(\cdot)$. We set $\gamma = (0.6, -0.6, 0.6, -0.6, 0, 0, \dots, 0)$ to denote four covariates (U_1, U_2, U_3, U_4) associated with assignment to treatment. Y_i are simulated from Bernoulli distributions with probability $Y_i = 1$ specified with $\Phi(\sum_{k=1}^p \phi_k U_{ik})$, with $\phi = (0.6, -0.6, 0, 0, 0.6, -0.6, 0, 0, \dots, 0)$ to denote four covariates (U_1, U_2, U_5, U_6) associated with Y . This implies a true value of $\Delta = 0$ and that the minimal model required to satisfy the assumption of strong ignorability is $\alpha^* = (1, 1, 0, 0, \dots, 0)$. All scenarios entail $p - 6$ extraneous covariates that are not associated with X or Y .

We simulate 1000 replicated datasets under each of several simulation scenarios. We analyze the simulated data with the three methods proposed in Section 3 and with three comparator methods: (a) a traditional sequential “kitchen sink” approach that includes all p variables in the PS model; (b) a forward “stepwise on X ” procedure that selects variables according to a Bayesian information criterion (BIC) applied to the PS model only, without acknowledging uncertainty in the variable selection; and (c) a “gold standard” approach that directly estimates the correct data-generating mechanism without making use of the PS. All Bayesian and approximately Bayesian inferences are based on MCMC chains run for 10,000 iterations, with the first 2000 discarded as burn-in and every 10th sample saved for posterior inference. MC³ analyses are run using R software (R Core Team 2013), SSVS analyses are programmed in C++, and all stated computing times are from a desktop computer running Mac OS X with a 2.66 GHz Intel Core i5 processor and 4 GB RAM. The gold standard, kitchen sink, and stepwise on X methods are fit using maximum likelihood, with the latter methods treating the estimated PS as fixed in the outcome stage. Careful analysis with the PS should generally involve checks of whether covariates are balanced between treated and untreated observations. We forego balance checks until the data analysis of Section 5.

4.1 Bayesian PS Model Averaging When Large p Requires Model Selection

We first consider scenarios where the kitchen sink approach is unavailable because p is too large relative to n , rendering some form of model selection necessary to estimate Δ with PS methods. We simulate two such scenarios: one with $n = 200, p = 100$ and another with $n = 500, p = 200$. Recall that the fully Bayesian MC³ approach requires the parametric assumptions outlined in Section 3.1, namely, that $h(\text{PS}(\gamma, \alpha, U_i); \xi)$ denotes linear adjustment for $\Phi^{-1}(\text{PS})$ and that residual adjustment is included for all except one variable having $\alpha_k = 1$ to prevent perfect collinearity. Also recall that these parametric assumptions reduce to a Probit regression model that includes only main

Table 1. Simulated scenarios where p is large relative to n . “FB MC3” refers to the fully Bayesian MC³ approach of Section 3.1, “AB MC3” refers to the approximately Bayesian MC³ approach of Section 3.2, and “FB SSVS” refers to the fully Bayesian SSVS approach of Section 3.3.

Table entries for the Bayesian procedures denote estimates of the posterior probability that U_k is included in the model ($p(\alpha_k = 1|\text{Data})$), averaged across replicated datasets with $n = 500$, $p = 200$ or $n = 200$, $p = 100$. Entries for the other approaches denote the proportion of 1000 replicated datasets for which U_k was selected. Values listed for $k > 6$ are the minimum and maximum value

	Covariate ID (k)							
	Assoc. with X and Y (α^*)		Assoc. with X only		Assoc. with Y only		Not associated	
	$k = 1$	2	3	4	5	6	$k > 6$ (min)	$k > 6$ (max)
$n = 200, p = 100$								
$n = 200$ FB MC3	1.000	1.000	0.982	0.984	0.985	0.983	0.021	0.057
$n = 200$ AB MC3	0.995	0.998	0.997	0.996	0.233	0.247	0.229	0.260
$n = 200$ FB SSVS	0.811	0.818	0.297	0.306	0.716	0.702	0.003	0.008
$n = 200$ stepwise on X	1.000	1.000	0.999	0.999	0.023	0.030	0.020	0.049
$n = 200$ kitchen sink	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 200$ gold	1.000	1.000	0.000	0.000	1.000	1.000	0.000	0.000
$n = 500, p = 200$								
$n = 500$ FB MC3	1.000	1.000	1.000	1.000	1.000	1.000	0.006	0.049
$n = 500$ AB MC3	1.000	1.000	1.000	1.000	0.084	0.087	0.073	0.093
$n = 500$ FB SSVS ⁺	1.000	1.000	0.992	0.992	1.000	1.000	0.001	0.006
$n = 500$ stepwise on X	1.000	1.000	1.000	1.000	0.018	0.019	0.008	0.025
$n = 500$ kitchen sink	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$ gold	1.000	1.000	0.000	0.000	1.000	1.000	0.000	0.000

NOTE: ⁺ Based on analysis of 340 datasets.

effects for selected variables and no PS, which, in this case, corresponds to the true data-generating mechanism. For the approximately Bayesian MC³, SSVS, kitchen sink, and stepwise on X approaches, we specify $h(\text{PS}(\gamma, U_i); \xi)$ to adjust for five subclasses defined by quintiles of the PS, and conduct residual adjustment for every variable included in the PS. We set $(\tau = 5, c = 200)$ for the prior specification in (8) and (9) for the SSVS approach. Computing time for the analysis of a single simulated dataset for the approximately Bayesian MC³ (SSVS) approach with $n = 200$, $p = 100$ was 58.7 (302.4) sec. Analogous values for the $n = 500$, $p = 200$ scenario were 102.9 (2169.7) sec. The computing time for the fully Bayesian MC³ approach was similar to that of the approximately Bayesian MC³ approach. The computational demand of the SSVS approach precluded the ability to analyze all 1000 replicated datasets for the $n = 500$, $p = 200$ scenario; results for the SSVS analysis of this scenario are based on analysis of 340 datasets.

Out of the 1000 datasets simulated under the scenario with $n = 200$, $p = 100$ ($n = 500$, $p = 200$), PS model estimates in 997 (916) were so unstable that estimated PS values were equal to 0 for all observations with $X_i = 0$ and equal to 1 for all observations with $X_i = 1$, meaning that any PS adjusted comparison is completely confounded and not interpretable as an estimate of the causal effect. Thus, in these settings, model selection is necessary to estimate causal effects with the PS. For the three proposed methods, Table 1 summarizes the average marginal probability that variable k is included in the PS model ($p(\alpha_k = 1|\text{Data})$) across the simulated datasets, along with the frequency with which each variable is included in the comparator methods. Figure 1 displays boxplots of posterior mean estimates from the Bayesian methods as well as point estimates from the comparator methods, along with bias and mean squared error

(MSE) for estimates from each method. Recall that the salient issue for estimating a causal effect is whether the method averages estimates across models including the variables that comprise α^* , here, (U_1, U_2) . From Table 1, we see that, in both scenarios, all three proposed methods average causal estimates across models that virtually always contain (U_1, U_2) , the notable exception coming for the SSVS approach with $n = 200$, $p = 100$, which produces average values of $p(\alpha_k = 1|\text{Data}) \approx 0.80$ for $k = 1, 2$. This implies that SSVS in this scenario (and with these values of τ, c) estimates Δ with weighted averages over models that frequently do not include all confounders, leading to the bias depicted in Figure 1(a). For the $n = 500$, $p = 200$ scenario, all procedures always include (U_1, U_2) , and exhibit negligible bias. In both scenarios, all methods tend to average estimates across models that include (U_3, U_4) that are only associated with X , with the SSVS analysis of the $n = 200$, $p = 100$ scenario including these variables substantially less often than the other methods. The distinction between the fully Bayesian approaches and the approximately Bayesian approach is highlighted in Table 1 by the average values of $p(\alpha_k = 1|\text{Data})$ for the variables only associated with Y (U_5, U_6); the SSVS and fully Bayesian MC³ approaches average causal estimates across models that tend to include these variables, whereas the approximately Bayesian MC³ approach does not. The variables (U_5, U_6) are included in the PS model least often with the stepwise on X procedure. The efficiency gain associated with the inclusion of (U_5, U_6) is highlighted in Figure 1, where the SSVS and fully Bayesian MC³ approach produce estimates with lower MSE than the approximately Bayesian or stepwise on X approaches, although recall that the fully Bayesian MC³ approach requires very restrictive parametric assumptions that happen to correspond to this simulated data generation. The fully Bayesian MC³ approach

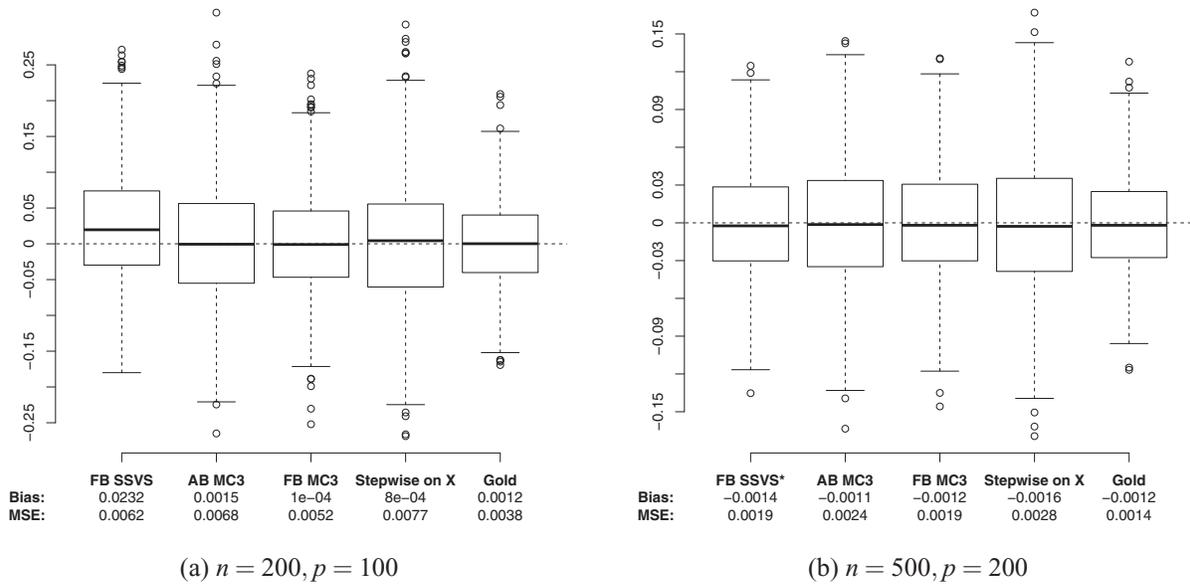


Figure 1. Simulated scenarios where p is large relative to n : boxplots of point estimates of Δ across 1000 simulated datasets with $n = 500, p = 200$ or $n = 200, p = 100$. “FB MC3” refers to the fully Bayesian MC³ approach of Section 3.1, “AB MC3” refers to the approximately Bayesian MC³ approach of Section 3.2, and “FB SSVS” refers to the fully Bayesian SSVS approach of Section 3.3. *SSVS analysis of $n = 500, p = 200$ scenario based on analysis of 340 datasets. (a) $n = 200, p = 100$, (b) $n = 500, p = 200$.

should not be expected to perform as well in practice. In summary, in a setting where the kitchen sink approach is unavailable and a choice regarding the PS model is required, the proposed methods can reliably select the variables in α^* and estimate the causal effect Δ with improved performance over a stepwise on X procedure that selects a single PS model without acknowledging the associated uncertainty.

4.2 Bayesian PS Model Averaging With Moderate p

In settings with more moderate p relative to n , the kitchen sink approach can be used to circumvent variable selection altogether. Here, we simulate three such settings to investigate whether our proposed methods can improve over the kitchen sink by averaging causal estimates across models that include only relevant variables. Toward this end, we simulate data with

sample sizes $n = 1000, 500$, and 200 , while holding p fixed at 20. For these simulations, we forego the restrictive fully Bayesian MC³ approach and analyze the simulated data with the approximately Bayesian MC³ approach of Section 3.2 and with the SSVS approach of Section 3.3, with $(\tau = 5, c = 200)$. For all methods in this section, we specify $h(\text{PS}(\gamma, U_i); \xi)$ to adjust for five subclasses defined by quintiles of the PS, and conduct residual adjustment with $\sum \alpha_k \delta_k U_{ik}$. Computing time for the analysis of a single simulated dataset for the approximately Bayesian MC³ (SSVS) approach was 100.6 (237.3), 69.3 (122.2), 51.2 (51.3) sec for the $n = 1000, 500, 200$ scenarios, respectively.

Figure 2 displays the average marginal inclusion probabilities, $p(\alpha_k = 1 | \text{Data})$ for $k = 1, 2, \dots, 20$, from the fully Bayesian SSVS analysis and the approximately Bayesian MC³ analysis across the 1000 replicated datasets with $n = 200, 500$,

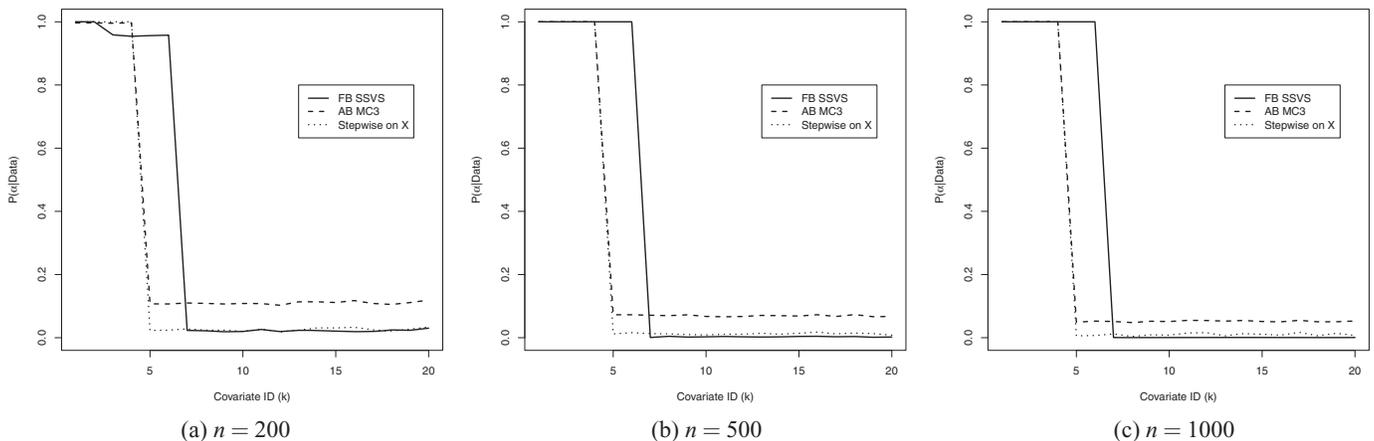


Figure 2. Simulated scenarios where $p = 20$ and $n = 200, 500$, or 1000 . Average values of $p(\alpha_k = 1 | \text{Data})$ for $k = 1, 2, \dots, 20$, averaged across 1000 simulated datasets. “FB SSVS” refers to the SSVS approach of Section 3.3, “AB MC3” refers to the approximately Bayesian approach of Section 3.2, and “Stepwise on X” refers to proportion of times each variable was selected with the forward BIC procedure described in Section 4. (a) $n = 200$, (b) $n = 500$, (c) $n = 1000$.

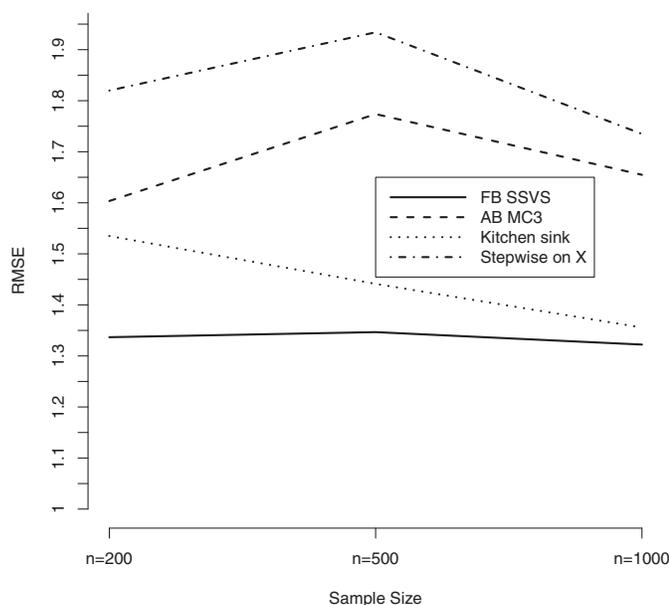


Figure 3. Simulated scenarios where $p = 20$ and $n = 200, 500,$ or 1000 : relative mean squared error (MSE) of estimates of Δ (relative to the “gold standard” approach), averaged across 1000 simulated datasets. “FB SSVS” refers to the SSVS approach of Section 3.3 and “AB MC3” refers to the approximately Bayesian approach of Section 3.2.

and 1000, compared with the proportion of times each variable was selected with the stepwise on X procedure. The Bayesian methods average causal estimates across models that always contain α^* , with $p(\alpha_k = 1 | \text{Data}) = 1.0$ for $k = 1, 2$ for all three scenarios, and the stepwise on X procedure always selects these variables. All methods virtually always include (U_3, U_4) associated with X only, with minimum average $p(\alpha_k = 1 | \text{Data}) = 0.95$ for $k = 3, 4$ from the Bayesian methods across the three scenarios. The major departure between the methods comes in the omission of the variables only associated with Y from PS models estimated with the approximately Bayesian MC³ approach and the stepwise on X approach, as compared with the SSVS approach that exhibits minimum average $p(\alpha_k = 1 | \text{Data}) = 0.96$ for $k = 5, 6$ across all three scenarios.

Figure 3 displays the relative MSE of posterior mean estimates of Δ and estimates from the stepwise on X and kitchen sink procedures, relative to estimates from the gold standard approach. All methods exhibit negligible bias (results not shown). Figure 3 illustrates that the SSVS approach produces estimates with MSE closest to the gold standard, and that the stepwise on X procedure yields the largest MSE. While the efficiency loss associated with including extraneous variables in the kitchen sink approach becomes more pronounced as the sample size decreases, the kitchen sink estimates still exhibit lower MSE than the approximately Bayesian MC³ approach that averages causal estimates across models that rarely include the variables associated with Y only (U_5, U_6).

Web Appendix D illustrates a simulation study with $n = 200$, $p = 20$ for which the U_k are correlated. Again, the SSVS approach produces estimates with MSE closest to the gold standard, and the stepwise on X approach exhibits the largest MSE. Unlike the uncorrelated data setting depicted in Figure 3, the ap-

proximately Bayesian approach produces estimates with lower MSE than the kitchen sink approach in the correlated data setting of Web Appendix D, and failure to always select the variables in α^* leads to small bias in estimates from the SSVS, approximately Bayesian MC³, and stepwise on X approaches.

These simulations illustrate that when p is moderate relative to n , the SSVS and approximately Bayesian MC³ methods for variable selection and model averaging can outperform (in terms of MSE) a stepwise on X selection procedure and a kitchen sink approach that circumvents variable selection entirely. Thus, we have shown that, in certain situations, the additional uncertainty associated with variable selection and model averaging can outweigh the drawbacks of a stepwise procedure or the inefficiency of extraneous adjustment that comes as a consequence of automatically including every potential confounder in the PS model.

5. COMPARING THE EFFECTIVENESS OF TREATMENTS FOR MALIGNANT BRAIN TUMORS

Standard treatment for malignant brain tumors (most commonly malignant astrocytomas such as glioblastoma) often involves surgical excision of the tumor, but recent evidence suggests that surgery is often foregone in elderly patients, especially those with poor preoperative status and of particularly advanced age (Iwamoto et al. 2008). Due in part to the relatively low incidence of these tumors, evidence as to whether elderly patients benefit from surgical treatment is unclear, and large administrative databases have emerged as a key tool for comparing the effectiveness of treatments for malignant brain tumors. We use data on 2606 Medicare beneficiaries residing in the Northeastern United States (Massachusetts, Connecticut, New Hampshire, Maine, Rhode Island, and Vermont) who were hospitalized with a primary diagnosis of malignant brain neoplasm during 2000–2009 to compare the effectiveness of surgical excision $X = 1$ versus no surgery $X = 0$ for lowering the risk of mortality within 1 year of diagnosis. To ensure that our study population did not include patients with brain metastases of other cancers, this patient sample excludes patients with a previous cancer diagnosis.

Table 2 summarizes the characteristics of patients who were treated with surgical excision ($n = 1118$) and those who did not receive surgery ($n = 1488$). As expected, patients aged over 85 years were much less likely to receive surgery, and patient characteristics measured by Hierarchical Condition Categories (Pope et al. 2004) capturing current or previous comorbidities were different in the two groups, with surgical patients exhibiting fewer comorbidities overall. An unadjusted comparison indicates that the rate of death within 1 year of hospitalization is 15.2% lower in surgical patients. We estimate the ACE of surgery (vs. no surgery) with the approximately Bayesian MC³ approach from Section 3.2, as well as with the fully Bayesian SSVS approach from Section 3.3 (with $\tau = 5$, $c = 1000$). For both methods, we specify $h(\text{PS}(\gamma, U_i); \xi)$ to adjust for five subclasses defined by quintiles of the PS, and conduct residual adjustment within PS subclass for every covariate included in the PS. As in our simulations, we specify Probit regression in the PS and outcome models. For both methods, MCMC chains were run for 150,000 iterations, discarding the first 50,000 as burn-in and saving every remaining 10th sample for posterior

Table 2. Baseline characteristics (% experiencing) and 1 year mortality rate for patients with and without surgery for brain tumor removal, along with estimated $p(\alpha_k = 1|Data)$ for the fully Bayesian SSVS and approximately Bayesian MC³ analyses of the Medicare data from the Northeastern US in 2000–2009

	No surgery (<i>n</i> = 1488)	Surgery (<i>n</i> = 1118)	$p(\alpha_k = 1 Data)$	
			SSVS	AB MC ³
Female	51.0	44.6	0.00	0.06
White	97.2	97.9	0.00	0.04
Congestive heart failure	5.0	2.6	0.00	0.10
Myocardial inarction	1.3	0.9	0.00	0.01
Chronic atherosclerosis	22.8	18.2	0.00	0.08
Respiratory failure	1.1	1.2	0.00	0.01
Valvular disease	6.5	5.4	0.00	0.01
Arrhythmia	7.2	4.7	0.03	0.02
Hypertension	63.2	60.5	0.00	0.01
Stroke	5.4	2.6	0.00	0.28
Cerebrovascular disease	4.0	2.5	0.00	0.04
Renal failure	2.7	1.2	0.11	0.05
COPD	10.0	10.7	0.00	0.02
Pneumonia	3.8	4.1	0.00	0.04
Diabetes	19.3	17.4	0.62	0.02
Dementia	15.2	9.5	0.33	0.68
Functional disability	4.0	1.8	0.00	0.25
Peripheral vascular disease	2.0	1.3	0.00	0.02
Trauma in the past year	5.0	3.6	0.00	0.02
Substance abuse	6.7	6.6	0.00	0.02
Major psychiatric disorder	5.4	3.2	0.00	0.35
Depression	8.5	7.2	0.00	0.02
Parkinson's/Huntington's	2.2	1.4	0.08	0.03
Seizure disorder	22.7	18.9	1.00	0.27
Chronic fibrosis	1.5	0.8	0.58	0.05
Asthma	2.9	3.3	0.01	0.02
Age 65–74	37.1	52.3	(reference)	
Age 75–84	43.1	41.9	1.00	0.99
Age 85+	19.8	5.8	1.00	1.00
Death within 1 Year	85.6	70.4		

inference. Computing times were 2965.9, 11712.4 sec for the approximately Bayesian MC³, SSVS procedure, respectively.

Covariate balance was assessed at each MCMC iteration by comparing covariate prevalence between surgical and nonsur-

gical patients within PS subclass, using only the covariates in the model at that iteration. Both methods similarly balanced all the covariates in Table 2 within PS subclass, with some age group imbalances persisting in the lowest PS subclass. Detailed balance checks appear in Web Appendix B.

The approximately Bayesian MC³ approach produced a posterior mean estimate of $\hat{\Delta} = -0.119$ and a 95% posterior interval $(-0.151, -0.087)$. Table 3 displays the 10 models with highest posterior support from this approach. Together, these 10 models comprise 34% of the total posterior model support, meaning that 66% is spread across models not appearing in Table 3. Every individual model not in Table 3 receives < 2% posterior support. Table 2 displays the marginal $p(\alpha_k = 1|Data)$ for $k = 1, \dots, 28$, across all 2²⁸ models. These tables indicate that the posterior estimate of Δ is a weighted average of estimates from models that virtually always include patient age group, but also commonly involve presence of stroke, dementia, functional disability, psychological disorder, and seizure disorder.

The fully Bayesian SSVS analysis produced a similar causal estimate, with a posterior mean $\hat{\Delta} = -0.122$, with 95% posterior probability interval $(-0.154, -0.090)$. Table 4 displays the 10 models with highest posterior support from the SSVS analysis. These 10 models comprise 97% of the posterior model support, and Table 2 displays the marginal $p(\alpha_k = 1|Data)$ for $k = 1, \dots, 28$. For the SSVS analysis, the posterior distribution of Δ is a weighted average of estimates from models that always include patient age group and seizure disorder, with substantial posterior support also distributed across models including renal failure, diabetes, dementia, and chronic fibrosis.

Recall that these two methods differ in two important ways and do not converge to the same stationary distribution. First, the approximately Bayesian MC³ approach builds PS models based only on covariate associations with X , whereas the fully Bayesian SSVS approach uses associations with both X and Y . Second, the approximately Bayesian MC³ approach includes variables on the basis of increases in the integrated likelihood term for the PS model, whereas SSVS includes variables based on the estimated magnitude of $(\gamma_k^2 + \delta_k^2)$ from (7) and the specification of hyperparameters (τ, c) . These differences are evident when comparing the results from these two methods. For

Table 3. Ten models with highest posterior support from the approximately Bayesian MC³ analysis of the Medicare data

Model <i>m</i>	α_k								$p(\alpha^m Data)$	$\hat{\Delta}^m$
	Stroke	Dementia	Functional Disability	Psych. Disorder	Seizure	Age 75–84	Age 85+			
1	0	1	0	0	0	1	1	0.08	-0.118	
2	1	1	0	0	0	1	1	0.05	-0.119	
3	0	1	1	0	0	1	1	0.05	-0.117	
4	0	1	0	0	1	1	1	0.03	-0.122	
5	0	1	0	1	0	1	1	0.03	-0.116	
6	1	1	0	1	0	1	1	0.02	-0.117	
7	0	0	1	0	0	1	1	0.02	-0.118	
8	0	1	1	1	0	1	1	0.02	-0.118	
9	0	1	0	1	1	1	1	0.02	-0.123	
10	1	0	0	1	0	1	1	0.02	-0.116	
Posterior mass of top 10 models:								0.34		

Table 4. Ten models with highest posterior support from the fully Bayesian SSVS analysis of the Medicare data

Model m	α_k										$p(\alpha^m \text{Data})$	$\hat{\Delta}^{\alpha^m}$
	Arrhythm.	Renal Fail	Diabetes	Dementia	Parkinson's/ Huntington's	Seizure	Chronic Fibrosis	Asthma	Age 75–84	Age 85+		
1	0	0	1	0	0	1	0	0	1	1	0.27	-0.122
2	0	0	0	1	0	1	1	0	1	1	0.24	-0.123
3	0	0	1	0	0	1	1	0	1	1	0.18	-0.122
4	0	0	1	0	1	1	1	0	1	1	0.08	-0.12
5	0	0	0	1	0	1	0	0	1	1	0.08	-0.124
6	0	1	0	0	0	1	1	0	1	1	0.05	-0.123
7	0	1	1	0	0	1	0	0	1	1	0.03	-0.121
8	1	0	1	0	0	1	0	0	1	1	0.02	-0.119
9	1	1	1	0	0	1	0	0	1	1	0.01	-0.12
10	0	1	0	0	0	1	1	1	1	1	0.01	-0.118
Posterior mass of top 10 models:											0.97	

example, estimates of Δ from the approximately Bayesian MC³ method virtually never rely on a model that includes diabetes, whereas diabetes appears in 68% of the models constructed with the fully Bayesian SSVS approach (see Tables 2–4). This likely occurs because diabetes is not predictive of having a surgical excision, but is strongly related to death within 1 year of hospitalization. However, comparison between the approximately and fully Bayesian procedures does not always follow so clearly. For example, history of stroke appears frequently in models estimated with the approximately Bayesian MC³ procedure ($p(\alpha_k = 1 | \text{Data}) = 0.28$), but does not appear in any model receiving posterior support from the SSVS procedure. Such an apparent discrepancy could arise for several reasons. It is possible that, after adjusting for variables that tend to appear in the SSVS analysis (e.g., chronic fibrosis), stroke is no longer related to X or Y . As a rough guide from which to judge variable associations with Y , Web Appendix C augments Table 2 with point estimates, standard errors, and p -values from a Probit regression model that does not include the PS but includes main effects for each of the 28 variables. Finally, it is important to note that the differences in implementation lead the MC³ approach to distribute support across a large set of models compared with the SSVS approach, which concentrates 79% of the posterior mass at the top three models in Table 4. The concentration of posterior mass on relatively few models with the SSVS approach is due to the feature that SSVS uses more information in the data (e.g., information pertaining to Y), and also to the specification of the hyper parameters (τ, c), which will generally affect which variables are selected and how often; for a fixed value of τ , larger values of c distribute more posterior support to models containing more covariates. The need to decide an appropriate value of c highlights one practical difficulty for implementing the SSVS approach, although results from analyses of these data with different values of c are not reported because causal estimates remained virtually unchanged.

For comparison, a standard sequential kitchen sink approach produced a point estimate of $\hat{\Delta} = -0.115$, with a 95% confidence interval of $(-0.146, -0.084)$ based on 1000 parametric bootstrap samples. In summary, all methods estimate a similar beneficial ACE of surgical treatment for preventing death within

1 year of hospitalization that is attenuated as compared with the simple unadjusted comparison.

6. DISCUSSION

For causal inference methods based on the PS, we have introduced Bayesian methods for variable selection and model averaging to estimate causal treatment effects that can offer improvement over standard PS approaches, especially when p is large and some form of variable selection is required. Rather than condition inference on a single PS model, our approaches average causal estimates across many different models that are supported by the observed data.

Aside from notions of model averaging and model uncertainty, our proposed fully Bayesian procedures entail another important distinction with traditional approaches that select variables based only on covariate associations with treatment assignment (X). The fully Bayesian approaches build PS models based in part on covariate associations with the outcome (Y) as well, which is important in light of the well-documented result that including outcome predictors in PS models can improve causal estimates (Rubin and Thomas 1996; Rubin 1997; Brookhart et al. 2006). This feature of our approach also relates to the notion of the “prognostic score” that balances covariates between treated and untreated units based *only* on associations with Y (Hansen 2008), although it remains unclear whether models for the prognostic score and the outcome, conditional on the score, could be combined in a joint likelihood for the type of Bayesian analysis pursued here.

Data-driven approaches to select important outcome predictors could be construed as in violation of one philosophical motivation of the PS. Specifically, Rubin (2008) argued that the PS is meant to approximate the design stage of a randomized study, and as such should be constructed without any access to the outcome to ensure objective design decisions that are completely separate from analysis decisions. It should be noted that the use of outcome information in our fully Bayesian approaches is completely automated, thus precluding the selection of variables based on post hoc knowledge of the estimated treatment effect. The approximately Bayesian strategy of Section 3.2

selects variables based only on strength of association with X , without regard to Y , and as such cannot be construed as violating the separation of design and analysis decisions. However, at least in our simulation studies, the approximately Bayesian approach did not perform as well (in terms of MSE) as the fully Bayesian approaches, nor did it perform as well as the kitchen sink approach when p was moderate and the covariates were uncorrelated. This highlights the potential benefit of including predictors of Y . One important limitation that our proposed methods share with nearly all methods for selecting variables to include in the PS is that they will include those that are strongly associated with X even if they are not at all associated with Y , and such variables have been shown to reduce efficiency of causal estimates (Rubin 1997; Brookhart et al. 2006). Rubin (1997) argued that the efficiency loss of including such variables should be relatively inconsequential if the variable is even weakly associated with Y . In general, relative performance of the proposed and traditional methods will depend on the nature of associations between covariates, treatment, and outcome.

One general feature of the proposed approaches is their reliance on parametric model specifications for the PS and outcome models. This was most pronounced in the fully Bayesian MC³ approach of Section 3.1, which motivated the more flexible approaches of Sections 3.2 and 3.3. Nonetheless, our proposed methods necessarily rely on model specification more than modern machine-learning approaches to building PS models, such as generalized boosted models (McCaffrey, Ridgeway, and Morral 2004), that can model the relationships between covariates and treatment assignment in a very flexible manner. In exchange for parametric specification, our methods have the benefit of selecting variables based in part on their association with Y , and also preclude the need to condition on a single PS model for inference. Efforts to improve the parametric flexibility of our methods are warranted, including exploration of approximations to the integrated likelihood that can circumvent the restrictive parametric assumptions required for a fully Bayesian MC³ analysis, possibly along the lines of the Bayes' factor approximations commonly used for generalized linear models (Hoeting et al. 1999). Alternatively, the methods described here could serve as a precursor step to define a reasonably sized set of candidate variables for use in model-averaged estimates that make use of more flexible parametric specifications.

We made use of two distinct computational strategies for Bayesian variable selection and model averaging, both having relative merits and limitations. The MC³ approach is computationally appealing because it relies on an expression of the likelihood with the (possibly high-dimensional) vector of model-specific coefficients integrated out, but required restrictive parametric assumptions for a fully Bayesian approach. In contrast, the SSVS approach accommodates general likelihood expressions, but shrinks variables close to zero rather than remove them from the model, entailing a heavy computational burden for large p . Furthermore, the shrinkage is governed by hyperparameters that must be specified, and different specifications can produce different results.

Modern decision making in the health sciences and other fields increasingly relies on very large observational databases containing a wealth of covariate information, highlighting the importance of issues regarding model uncertainty in PS meth-

ods. This is especially true in comparative effectiveness investigations leveraging information in administrative databases, but could also prove integral in other “big data” applications that routinely deal with very large numbers of potential confounding factors. However, we stress that our data-driven approaches for variable selection and model averaging cannot replace careful scientific thought. For example, we considered situations where all U were known to be pretreatment covariates, but, in general, particular attention should be paid to whether some available variables are affected by treatment, leaving them inappropriate for PS adjustment. Finally, our methods do not provide a means for adjusting for unmeasured confounding; we assume that all confounders necessary to satisfy the assumption of ignorable treatment assignment are an unknown subset of those available for analysis. This is especially relevant in the results of our data analysis, where the Medicare data lack disease-specific information, thus presenting the possibility of unmeasured confounding.

APPENDIX A: DETAILS OF THE MC³ APPROACHES FOR PROPENSITY SCORE VARIABLE SELECTION AND MODEL AVERAGING

A.1 Prior Distributions

The MC³ approaches of Sections 3.1 and 3.2 rely on prior specification for $p(\beta, \xi, \gamma, \delta, \alpha)$, which we factorize as $p(\alpha)p(\gamma_\alpha|\alpha)p(\delta_\alpha|\alpha)p(\beta, \xi)$, where $(\gamma_\alpha, \delta_\alpha)$ denotes the components of (γ, δ) included in model α . As noted in Section 3, we assume equal prior probability for all possible α , $p(\alpha) = \frac{1}{M}$. For $p(\gamma_\alpha|\alpha)$, we assume a Normal distribution with mean 0 and variance $\lambda_\gamma(\mathbf{U}'_\alpha \mathbf{U}_\alpha)^{-1}$. For $p(\delta_\alpha|\alpha)$, we use a Normal distribution with mean 0 and variance $\lambda_\delta(\mathbf{U}'_{\alpha,Y} \mathbf{U}_{\alpha,Y})^{-1}$ for the fully Bayesian MC³ approach, and a Normal distribution with mean 0 and variance $10^2 \mathbf{I}$ for the approximately Bayesian approach, where \mathbf{I} is the identity matrix. Here \mathbf{U}_α represents the subset of \mathbf{U} contained in α (with dimension $n \times p_\alpha$), and $\mathbf{U}_{\alpha,Y}$ denotes the subset of \mathbf{U}_α included for residual adjustment in the outcome model (with dimension $n \times p_{\alpha,Y}$). We henceforth consider $\lambda_\gamma = \lambda_\delta = \lambda$. All analyses of simulated and Medicare data assume $\lambda = 4n$. Furthermore, we assume $p(\beta, \xi)$ follows a Normal distribution with mean 0 and variance $10^2 \mathbf{I}$.

A.2 MCMC Algorithm for the Fully Bayesian MC³ Approach

Variable selection with MC³ relies on an expression for $\frac{p(\alpha'|\text{Data})}{p(\alpha|\text{Data})}$ that governs the probability of the MCMC chain transitioning from α to a neighboring model α' . For the fully Bayesian approach of Section 3.1, the integrated likelihood can be expressed as $\int \int L(\mathbf{Y}, \mathbf{X}|\mathbf{U}, \gamma, \alpha, \beta, \delta)p(\gamma|\alpha)p(\delta|\alpha)p(\alpha)p(\beta, \xi)d\gamma d\delta$, with $L(\mathbf{Y}, \mathbf{X}|\mathbf{U}, \gamma, \alpha, \beta, \delta)$ as in (3) and (4). Note that the resulting expression will also depend on ξ and β , as the appearance of these parameters does not depend on α , but we omit this conditioning to simplify notation. Similarly, we omit the prior distribution $p(\beta, \xi)$ from the following expressions.

We follow Albert and Chib (1993) and specify a Probit link for both $g_x(\cdot)$ and $g_y(\cdot)$, which allows an MCMC data-augmentation procedure that iteratively samples Normally distributed latent continuous data with unit variance such that the latent $X^*(Y^*)$ are > 0 when $X = 1(Y = 1)$, and < 0 otherwise. This is Assumption (i) of Section 3.1. Conditional on simulated $(\mathbf{X}^*, \mathbf{Y}^*)$, the integrated likelihood can be

expressed as

$$\int \int (2\pi)^{-0.5(n+p_\alpha+n+p_{\alpha,Y})} \det[\lambda(\mathbf{U}'_\alpha \mathbf{U}_\alpha)^{-1}]^{-0.5} \det[\lambda(\mathbf{U}'_{\alpha,Y} \mathbf{U}_{\alpha,Y})^{-1}]^{-0.5} \times \exp \left\{ -0.5 \left[(\tilde{\mathbf{X}}^* \tilde{\mathbf{X}}^*) + (\tilde{\mathbf{Y}}^* \tilde{\mathbf{Y}}^*) + \gamma'_\alpha \left(\frac{1}{\lambda} \mathbf{U}'_\alpha \mathbf{U}_\alpha \right) \gamma_\alpha + \delta'_\alpha \left(\frac{1}{\lambda} \mathbf{U}'_{\alpha,Y} \mathbf{U}_{\alpha,Y} \right) \delta_\alpha \right] \right\} \mathbf{d}\gamma \mathbf{d}\delta \tag{A.1}$$

where $\tilde{\mathbf{X}}^* = (\mathbf{X}^* - \mathbf{U}_\alpha \gamma_\alpha)$ and $\tilde{\mathbf{Y}}^* = (\mathbf{Y}^* - \beta_0 \mathbf{1} - \beta_X \mathbf{X} - \mathbf{h}(\text{PS}(\gamma, \alpha, \mathbf{U}); \xi) - \mathbf{U}_{\alpha,Y} \delta)$. Note that integration with respect to γ in (A.1) is relatively straightforward if $\mathbf{h}(\text{PS}(\gamma, \alpha, \mathbf{U}))$ is chosen such that the term in $\exp\{\cdot\}$ remains quadratic in γ , because this will allow simplifications relying on the Normal probability distribution function. If the term in $\exp\{\cdot\}$ is not quadratic in γ , closed-form integration in (A.1) will prove problematic. This motivates Assumptions (ii) and (iii) in Section 3.1; specifying $h(\text{PS}(\gamma, \alpha, U_i); \xi) = \xi \times \sum \alpha_k \gamma_k U_{ik}$ to denote an outcome model that adjusts for $\Phi^{-1}(\text{PS}(\gamma, \alpha, U_i)) = \sum \alpha_k \gamma_k U_{ik}$ as a linear covariate facilitates closed-form evaluation of (A.1) with respect to γ , and one covariate in α must be removed from the residual adjustment term $\sum \alpha_k \delta_k U_{ik}$ to prevent perfect collinearity with the $\sum \alpha_k \gamma_k U_{ik}$. This strategy implies that if \mathbf{U}_α is of dimension $n \times p_\alpha$, then $\mathbf{U}_{\alpha,Y}$ will be the same except for the removal of the intercept term and one covariate, resulting in $p_{\alpha,Y} = p_\alpha - 2$.

Accordingly, substituting $\mathbf{h}(\text{PS}(\gamma, \alpha, \mathbf{U}))$ with $\xi \times (\mathbf{U}_\alpha \gamma_\alpha)$ in (A.1) and integrating over γ and δ yields

$$p(\alpha|\text{Data}) \propto (2\pi)^{-n} (1 + \lambda \xi^2 + \lambda)^{-\frac{p_\alpha}{2}} \lambda^{-\frac{p_{\alpha,Y}}{2}} \det[\mathbf{V}_\delta]^{-\frac{1}{2}} \det[\mathbf{U}'_{\alpha,Y} \mathbf{U}_{\alpha,Y}]^{-\frac{1}{2}} \times \exp \left\{ -0.5 [\mathbf{X}^* \mathbf{X}^* + \tilde{\mathbf{W}}' \tilde{\mathbf{W}} - (\mathbf{X}^* + \xi \tilde{\mathbf{W}})' \mathbf{U}_\alpha \mathbf{V}_\gamma \mathbf{U}'_\alpha + (\mathbf{X}^* + \xi \tilde{\mathbf{W}}) - \tilde{\mathbf{D}}' \mathbf{U}_\alpha \mathbf{V}_\delta \mathbf{U}'_\alpha \tilde{\mathbf{D}}] \right\}, \tag{A.2}$$

where $\tilde{\mathbf{W}} = (\mathbf{Y}^* - \beta_0 \mathbf{1} - \beta_X \mathbf{X})$, $\tilde{\mathbf{D}} = (\mathbf{U}'_{\alpha,Y} \tilde{\mathbf{W}} - \xi \mathbf{U}'_{\alpha,Y} \mathbf{U}_\alpha \mathbf{V}_\gamma \mathbf{U}'_\alpha (\mathbf{X}^* + \xi \tilde{\mathbf{W}}))$, $\mathbf{V}_\gamma^{-1} = (\mathbf{U}'_\alpha \mathbf{U}_\alpha (\frac{1}{\lambda} + \xi^2 + 1))$, and $\mathbf{V}_\delta^{-1} = (\mathbf{U}'_{\alpha,Y} \mathbf{U}_{\alpha,Y} (1 + \frac{1}{\lambda}) - \mathbf{U}'_{\alpha,Y} \mathbf{U}_\alpha \mathbf{V}_\gamma \mathbf{U}'_\alpha \mathbf{U}_{\alpha,Y} \xi^2)$.

The MC³ algorithm described in Section 3.1 is outlined as follows for iteration $(t + 1)$, considering current values of $\alpha^{(t)}$, $\gamma^{(t)}$, $\delta^{(t)}$, $\beta^{(t)}$, $\xi^{(t)}$, $\mathbf{X}^{*(t)}$, $\mathbf{Y}^{*(t)}$:

1. Propose α' by adding or removing one term from $\alpha^{(t)}$.
2. Set $\alpha^{(t+1)} = \alpha'$ with probability $\min\{1, \frac{p(\alpha'|\text{Data})}{p(\alpha^{(t)}|\text{Data})}\}$, where $p(\alpha|\text{Data})$ is as in (A.2). Set all elements of γ and δ not included in $\alpha^{(t+1)}$ to 0.
3. Update $(\beta^{(t+1)}, \xi^{(t+1)})$ from random-walk Metropolis step with values proposed from a Normal distribution with mean $(\beta^{(t)}, \xi^{(t)})$ and proposal variance $0.1\mathbf{I}$.
4. Update $\delta_\alpha^{(t+1)}$ from $N(\mathbf{M}_\delta, \mathbf{V}_\delta)$, where $\mathbf{M}_\delta = \mathbf{V}_\delta \tilde{\mathbf{D}}$.
5. Update $\gamma_\alpha^{(t+1)}$ from $N(\mathbf{M}_\gamma, \mathbf{V}_\gamma)$, where $\mathbf{M}_\gamma = \mathbf{V}_\gamma (\mathbf{U}'_\alpha (\mathbf{X}^* + \xi \tilde{\mathbf{W}} - \xi \mathbf{U}_{\alpha,Y} \delta))$ and recalculate the corresponding PS, $\mathbf{U}_\alpha \gamma_\alpha$.
6. Simulate new \mathbf{X}^* from a truncated Normal distribution with mean $\mathbf{U}_\alpha \gamma_\alpha$ and variance 1, and simulate new \mathbf{Y}^* from a truncated normal distribution with mean $\beta_0 \mathbf{1} + \beta_X \mathbf{X} + \xi \mathbf{U}_\alpha \gamma_\alpha + \mathbf{U}_{\alpha,Y} \delta$ and variance 1, as in Albert and Chib (1993).

A.3 Implied Parameterization of Outcome Surface for the Fully Bayesian MC³ Approach

Without loss of generality, consider $\alpha = (1, 1, \dots, 1)$ to denote inclusion all p available covariates. With assumptions (ii) and (iii), the linear term in the outcome model (2) can be expressed as

$$\beta_0 + \beta_X X + \xi(\gamma_0 + \gamma_1 U_1 + \gamma_2 U_2 + \dots + \gamma_p X_p) + \delta_2 U_2 + \delta_3 U_3 + \dots + \delta_p U_p, \tag{A.3}$$

where, again without loss of generality, U_1 is removed from the residual adjustment in accordance with assumption (iii). Expression (A.3) can be rewritten as $\beta_0^* + \beta_X X + \xi \gamma_1 U_1 + \beta_2^* U_2 + \beta_3^* U_3 + \dots + \beta_p^* U_p$,

where $\beta_0^* = (\beta_0 + \xi \gamma_0)$, and $\beta_k^* = (\xi \gamma_k + \delta_k)$ for $k = 2, 3, \dots, p$. This corresponds to an outcome model that is simply a reparameterization of a model entailing an additive main effect term for every covariate, which is precisely the type of parametric outcome model PS methods are meant to avoid.

A.4 MCMC Algorithm for the Approximately Bayesian MC³ Approach

As laid out in Section 3.2, an approximately Bayesian approach follows from adapting the above strategy to update (α, γ) conditional on quantities that appear only in the PS model, then update parameters (β, ξ, δ) conditional on (α, γ) . This simplifies the expression for $\frac{p(\alpha'|\text{Data})}{p(\alpha|\text{Data})}$ because of the simplified form of the integrated likelihood:

$$\int (2\pi)^{-\frac{(n+p_\alpha)}{2}} \det[\lambda(\mathbf{U}'_\alpha \mathbf{U}_\alpha)^{-1}]^{-\frac{1}{2}} \times \exp \left\{ -0.5 \left[(\tilde{\mathbf{X}}^* \tilde{\mathbf{X}}^*) + \gamma'_\alpha \left(\frac{1}{\lambda} \mathbf{U}'_\alpha \mathbf{U}_\alpha \right) \gamma_\alpha \right] \right\} \mathbf{d}\gamma, \tag{A.4}$$

$$= (2\pi)^{-\frac{n}{2}} (\lambda + 1)^{-\frac{p_\alpha}{2}} \times \exp \left\{ -0.5 \left[\mathbf{X}^* \mathbf{X} - \mathbf{X}^* \mathbf{U}_\alpha \left(\left(1 + \frac{1}{\lambda} \right) \mathbf{U}'_\alpha \mathbf{U}_\alpha \right)^{-1} \mathbf{U}'_\alpha \mathbf{X}^* \right] \right\}$$

evaluation of which does not depend on $h(\text{PS}(\gamma, \alpha, U_i); \xi)$, β , δ , or Y . Thus, the approximately Bayesian MC³ approach is outlined as follows:

1. Propose α' by adding or removing one term from $\alpha^{(t)}$.
2. Set $\alpha^{(t+1)} = \alpha'$ with probability $\min\{1, \frac{p(\alpha'|\text{Data})}{p(\alpha^{(t)}|\text{Data})}\}$, where $p(\alpha|\text{Data})$ is as in (A.4). Set all elements of γ and δ not included in $\alpha^{(t+1)}$ to 0.
3. Update $\gamma_\alpha^{(t+1)}$ from $N(\frac{\lambda}{1+\lambda} (\mathbf{U}'_\alpha \mathbf{U}_\alpha)^{-1} \mathbf{U}'_\alpha \mathbf{X}^*, \frac{\lambda}{1+\lambda} (\mathbf{U}'_\alpha \mathbf{U}_\alpha)^{-1})$.
4. Update $(\beta, \xi, \delta_\alpha)^{(t+1)}$ from $N((\tilde{\mathbf{Z}}' \tilde{\mathbf{Z}} + (0.1)^2 \mathbf{I})^{-1} (\tilde{\mathbf{Z}}' \mathbf{Y}^*), (\tilde{\mathbf{Z}}' \tilde{\mathbf{Z}} + (0.1)^2 \mathbf{I})^{-1})$, where $\tilde{\mathbf{Z}}$ is the design matrix with columns representing an intercept, \mathbf{X} , $\mathbf{h}(\text{PS}(\gamma, \alpha, \mathbf{U}))$, and $\mathbf{U}_{\alpha,Y}$.
5. Simulate new \mathbf{X}^* from a truncated Normal distribution with mean $\mathbf{U}_\alpha \gamma_\alpha$ and variance 1, and simulate new \mathbf{Y}^* from a truncated normal distribution with mean $\beta_0 \mathbf{1} + \beta_X \mathbf{X} + \xi \mathbf{h}(\text{PS}(\gamma, \alpha, \mathbf{U})) + \mathbf{U}_{\alpha,Y} \delta$ and variance 1, as in Albert and Chib (1993).

SUPPLEMENTARY MATERIALS

Web Appendix: A supplementary appendix with additional simulation studies, a complement to the analysis of Medicare data of Section 5, and a detailed account of the MCMC algorithm for the SSVS approach of Section 3.3.

[Received October 2012. Revised October 2013.]

REFERENCES

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679. [105,106]
 Austin, P. C. (2008), "A Critical Appraisal of Propensity-Score Matching in the Medical Literature Between 1996 and 2003," *Statistics in Medicine*, 27, 2037–2049. [95]
 Bang, H., and Robins, J. M. (2005), "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–973. [96]

- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006), "Variable Selection for Propensity Score Models," *American Journal of Epidemiology*, 163, 1149–1156. [95,104]
- Gelman, A., and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Vol. 648), New York: Cambridge University Press. [97]
- George, E. I., and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–374. [98,99]
- Hansen, B. B. (2008), "The Prognostic Analogue of the Propensity Score," *Biometrika*, 95, 481–488. [104]
- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–401. [97,98,105]
- Iwamoto, F. M., Reiner, A. S., Panageas, K. S., Elkin, E. B., and Abrey, L. E. (2008), "Patterns of Care in Elderly Glioblastoma Patients," *Annals of Neurology*, 64, 628–634. [102]
- Little, R., and An, H. (2004), "Robust Likelihood-Based Analysis of Multivariate Data With Missing Values," *Statistica Sinica*, 14, 949–968. [96]
- Liu, F., Bayarri, M. J., and Berger, J. O. (2009), "Modularization in Bayesian Analysis, With Emphasis on Analysis of Computer Models," *Bayesian Analysis*, 4, 119–150. [97,98]
- Lunn, D., Best, N., Spiegelhalter, D., Graham, G., and Neuenschwander, B. (2009), "Combining MCMC With 'Sequential' PKPD Modelling," *Journal of Pharmacokinetics and Pharmacodynamics*, 36, 19–38. [97,98]
- Madigan, D., York, J., and Allard, D. (1995), "Bayesian Graphical Models for Discrete Data," *International Statistical Review/Revue Internationale de Statistique*, 63, 215–232. [98]
- McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004), "Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies," *Psychological Methods*, 9, 403–425. [105]
- McCandless, L. C. (2012), "Discussion of Adjustment Uncertainty and Propensity Scores," *Biometrics*, 68, 678–680. [96]
- McCandless, L. C., Gustafson, P., and Austin, P. C. (2009), "Bayesian Propensity Score Analysis for Observational Data," *Statistics in Medicine*, 28, 94–112. [96,97]
- Pope, G. C., Kautter, J., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Lezzoni, L. I., Ingber, M. J., Levy, J. M., and Robst, J. (2004), "Risk Adjustment of Medicare Capitation Payments Using the CMS-HCC Model," *Health Care Financing Review*, 25, 119–141. [102]
- Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191. [98]
- R Core Team (2013), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [99]
- Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [95,96]
- Rubin, D. B. (1997), "Estimating Causal Effects From Large Data Sets Using Propensity Scores," *Annals of Internal Medicine*, 127, 757–763. [95,104]
- (2008), "For Objective Causal Inference, Design Trumps Analysis," *The Annals of Applied Statistics*, 2, 808–840. [95,104]
- Rubin, D. B., and Thomas, N. (1996), "Matching Using Estimated Propensity Scores: Relating Theory to Practice," *Biometrics*, 52, 249–264. [95,104]
- Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009), "High-Dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data," *Epidemiology*, 20, 512–522. [95]
- Vansteelandt, S. (2012), "Discussions," *Biometrics*, 68, 675–678. [96]
- Wang, C., Parmigiani, G., and Dominici, F. (2012), "Bayesian Effect Estimation Accounting for Adjustment Uncertainty," *Biometrics*, 68, 661–671. [96,98]
- Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013), "Model Feedback in Bayesian Propensity Score Estimation," *Biometrics*, 69, 263–273. [96,97]