

Bayesian propensity score analysis for observational data

Lawrence C. McCandless^{1,*,\dagger,\ddagger}, Paul Gustafson² and Peter C. Austin^{3,4,5}

¹*Faculty of Health Sciences, Simon Fraser University, Canada*

²*Department of Statistics, University of British Columbia, Canada*

³*Department of Public Health Sciences, University of Toronto, Canada*

⁴*Department of Health Policy, Management and Evaluation, University of Toronto, Canada*

⁵*Institute for Clinical and Evaluative Sciences, Canada*

SUMMARY

In the analysis of observational data, stratifying patients on the estimated propensity scores reduces confounding from measured variables. Confidence intervals for the treatment effect are typically calculated without acknowledging uncertainty in the estimated propensity scores, and intuitively this may yield inferences, which are falsely precise. In this paper, we describe a Bayesian method that models the propensity score as a latent variable. We consider observational studies with a dichotomous treatment, dichotomous outcome, and measured confounders where the log odds ratio is the measure of effect. Markov chain Monte Carlo is used for posterior simulation. We study the impact of modelling uncertainty in the propensity scores in a case study investigating the effect of statin therapy on mortality in Ontario patients discharged from hospital following acute myocardial infarction. Our analysis reveals that the Bayesian credible interval for the treatment effect is 10 per cent wider compared with a conventional propensity score analysis. Using simulations, we show that when the association between treatment and confounders is weak, then this increases uncertainty in the estimated propensity scores. Bayesian interval estimates for the treatment effect are longer on average, though there is little improvement in coverage probability. A novel feature of the proposed method is that it fits models for the treatment and outcome simultaneously rather than one at a time. The method uses the outcome variable to inform the fit of the propensity model. We explore the performance of the estimated propensity scores using cross-validation. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: confounding; bias; observational studies; causal inference; Bayesian statistics

*Correspondence to: Lawrence C. McCandless, Faculty of Health Sciences, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A 1S6.

†E-mail: mccandless@sfu.ca

‡Assistant Professor of Biostatistics.

Contract/grant sponsor: Canadian Institutes of Health Research Team Grant in Cardiovascular Outcomes Research
Contract/grant sponsor: Heart and Stroke Foundation of Ontario; contract/grant number: NA 5703
Contract/grant sponsor: British Columbia Michael Smith Foundation for Health Research

1. INTRODUCTION

Analytic adjustment techniques using treatment propensity scores are popular for reducing confounding bias in observational studies of the effects of drug therapies [1–3]. The idea is to use the propensity score, defined as the probability of treatment given measured confounders, as a tool to ensure similarity in treated and untreated groups with respect to outcome risk factors [4, 5]. Patients with the same propensity score have the same distribution of measured confounders. Comparing treatment groups, conditional on the propensity score, yields unconfounded estimates of the treatment effect. To adjust for confounding, we can include the propensity score as a covariate in a regression model for the outcome variable. Rosenbaum and Rubin [5] recommend stratifying patients on quintiles of the propensity score. Other methods using regression and weighting are also available [1, 6].

Because the propensity score is unknown, the application of propensity score techniques involves a two-step procedure. First, the propensity score is estimated for each patient, typically from the fitted values of logistic regression of the treatment variable on measured confounders. Then, patients are stratified on quintiles of the estimated propensity scores. Standard errors for the treatment effect estimate are usually calculated without acknowledging uncertainty in the estimated propensity scores [1, 6]. Intuitively, confidence intervals for the treatment effect may be falsely precise.

There is little previous research investigating the impact of modelling uncertainty in the propensity score on treatment effect estimation. Tu and Zhou [7] propose an approach to interval estimation using the bootstrap. They show that their method increases the variance of treatment effect estimates, but do not consider a full comparison with conventional interval estimation procedures, which ignore uncertainty in the propensity score. A similar estimation strategy is proposed by Hirano and Imbens [8]. In contrast, there is a large body of work investigating the merits of using estimated propensity scores rather than true propensity scores (see Rubin [2] and the references therein). Other recent work on propensity score methods considers treatment effect estimation in settings where covariates are omitted from the propensity score model [9–12].

Bayesian methods offer a natural strategy for modelling uncertainty in the propensity scores. We can model the joint distribution of the data and parameters with the propensity score as a latent variable. Markov chain Monte Carlo (MCMC) methods allow simulation from the posterior distribution for model parameters. The marginal posterior for the treatment effect incorporates uncertainty in the propensity scores because it integrates over the latent variable. An advantage of this approach is that it permits the incorporation of Bayesian machinery into data analysis applications using propensity scores [13]. We may include prior information arising from expert opinion or previous research. Complex modelling assumptions for hierarchical data, measurement error or missing data can be incorporated.

Nonetheless, there are arguments against combining Bayesian analysis with propensity score techniques. It has been argued that the conditional distribution of the outcome within treatment groups should not depend on measured confounders only through the propensity score [13–15]. Propensity scores model information about the manner in which the study is designed and should convey no information about treatment effects. Other authors contend that including propensity scores in a Bayesian analysis can ease model specification and yield estimates with good frequentist properties [14, 16].

Given the speculative pros and cons, the objective of the present paper is to investigate the impact of modelling uncertainty in the estimated propensity scores using Bayesian techniques. We propose a Bayesian propensity score analysis (BPSA) for the case of stratifying on five subclasses of the

propensity score. We consider observational studies with a dichotomous treatment, dichotomous outcome, and measured confounders where the log odds ratio is the measure of effect. As a case study, Section 1.1 introduces an observational study of the effectiveness of statin therapy in patients discharged from Ontario hospitals following acute myocardial infarction. We use the data of Austin and Mamdani [1] who conducted a detailed comparison of propensity score methods. In Section 2, we outline our proposed BPSA including the model, prior distributions, and a method for posterior simulation using MCMC. In Section 3, we apply BPSA to the case study and study the impact of modelling uncertainty in the propensity scores. Our analysis reveals that the Bayesian credible interval for the treatment effect is 10 per cent wider compared with a conventional propensity score analysis. To study the performance of Bayesian interval estimates in more generality, Section 4 presents a simulation study where data are generated under competing models for the outcome. We show that when association between the treatment and confounders is weak, then this increases uncertainty in the estimated propensity scores. Bayesian interval estimates for the treatment effect are longer on average, though there is little improvement in coverage probability. A feature of BPSA is that it fits models for the treatment and outcome simultaneously rather than one at a time. The method uses the outcome variable in order to inform the fit of the propensity model. While this approach to estimation emerges naturally from Bayesian latent variable modelling, it is also unusual. To explore the performance of the estimated propensity scores from BPSA, Section 5 investigates prediction error using cross-validation techniques. Section 6 summarizes the paper and discusses the competing perspectives on the role of the outcome variable in propensity score modelling.

1.1. A motivating example: the EFFECT data

To motivate our methodology, we consider the example of an observational study estimating the effect of statin therapy, a class of lipid-lowering medications, on all-cause mortality in Ontario residents following hospitalization for acute myocardial infarction [1]. Detailed clinical data were obtained for 4572 patients discharged from Ontario hospitals between 1999 and 2000. For each patient, medical charts were abstracted to obtain information on demographic characteristics, cardiac risk factors, comorbid conditions, vascular history, vital signs at hospital admission, and laboratory tests. Records of medication prescriptions were also collected. The data were obtained in conjunction with the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study, an ongoing initiative to improve the quality of health services for Ontario residents with cardiovascular disease [1]. Patients were classified as statin users if they were prescribed a statin at hospital discharge and as statin non-users otherwise. Death within 3 years of hospital discharge was established by linking patient records to the Ontario Registered Persons Database.

Table I summarizes the characteristics of patients at hospital discharge and illustrates that the crude association between statin therapy and mortality is likely to be confounded. Treated patients are generally younger and healthier than untreated patients. This finding is consistent with previous studies of physician-prescribing habits, which show that statins are underprescribed in elderly patients with poor prognosis [17, 18]. Because clinical studies have shown that statin therapy reduces the risk of cardiovascular disease [19], we expect that the protective effect of statin therapy will be exaggerated in a comparison of mortality rates in treated versus untreated patients. Indeed, the crude odds ratio for the association between statin therapy and mortality is equal to 0.50 with 95 per cent confidence interval (0.42, 0.60), which is lower than the previously published estimates [19]. Analytic adjustment for confounding is required in order to estimate the treatment effect.

Table I. Baseline characteristics of 4572 patients discharged from hospital following acute myocardial infarction.

Characteristic	Statin prescribed (<i>n</i> = 1354)	Statin not prescribed (<i>n</i> = 3218)
	Number (per cent) or mean \pm SD	
<i>Demographic characteristics</i>		
Age (years)	63 \pm 12	68 \pm 14
Female sex	398 (29)	1201 (37)
<i>Presenting characteristics</i>		
Shock	\leq 5 (\leq 1)	24 (1)
<i>AMI risk factors</i>		
Family history of CAD	525 (39)	973 (30)
Diabetes	459 (26)	1060 (26)
CVA/TIA	122 (9)	312 (10)
High BP	548 (48)	1386 (43)
Current smoker	459 (34)	1060 (33)
Hyperlipidaemia	794 (59)	604 (19)
<i>Comorbidities</i>		
Angina	504 (37)	999 (31)
Renal disease	10 (1)	13 (\leq 1)
<i>Vital signs on admission</i>		
Systolic BP (mmHg)	149 \pm 31	148 \pm 32
Diastolic BP (mmHg)	85 \pm 18	84 \pm 18
Heart rate (beats/min)	81 \pm 23	84 \pm 23
Respiratory rate (beats/min)	20 \pm 5	21 \pm 6
<i>Laboratory values</i>		
White blood count (cell 10^9 /L)	10 \pm 5	10 \pm 5
Haemoglobin (g/L)	141 \pm 17	137 \pm 19
Sodium (mmol/L)	139 \pm 3	139 \pm 4
Glucose (mmol/L)	9 \pm 6	9 \pm 5
Creatinine (μ mol/L)	101 \pm 54	104 \pm 60

Austin and Mamdani [1] analyze the EFFECT data and argue that regression modelling of mortality risk is challenging. There are numerous prognostic variables and their relationship with mortality is poorly understood. In contrast, statin prescribing can be modelled more easily through consultation with physicians. The authors estimate the patient propensity scores and use the quantities to control confounding using a variety of methods including covariate adjustment, weighting, and matching. None of the methods model uncertainty in the estimated propensity scores and it is unknown whether this affects estimation.

2. BAYESIAN PROPENSITY SCORE ANALYSIS (BPSA) FOR OBSERVATIONAL DATA

We outline a method for BPSA, which uses two regression models; one for treatment assignment and one for the outcome. Rather than substituting propensity score estimates into the model for mortality risk, they are modelled as latent variables, which are integrated from the posterior distribution using MCMC.

2.1. Data, model and parameters

For the EFFECT data with sample size $n = 5472$, let X_i for $i = 1, \dots, n$ denote dichotomous random variables taking values 1 or 0 to model whether or not the i th patient was prescribed a statin at hospital discharge. Let $\{Y(1)_i, Y(0)_i\}$ denote dichotomous variables taking values 1 or 0 to model the potential outcomes for death in the i th patient [6]. Let C_i be a $p \times 1$ vector of measured confounders where the first component is equal to 1 to ease specification of y-intercept terms in regression modelling. Finally, let $Y_i = Y(X_i)_i$ denote the observed outcome and let $Z_i = \Pr(X_i = 1 | C_i)$ denote the propensity score for the i th patient.

Rosenbaum and Rubin [5] recommend stratifying on five subclasses of the estimated propensity scores. Omitting the subscript i from (Y_i, X_i, C_i) , we write the probability density function of Y and X given C as

$$p(Y, X|C) = p(Y|X, C)p(X|C)$$

For our proposed BPSA, we use two logistic regression models

$$\text{logit}[\Pr(Y = 1|X, C)] = \beta X + \xi^T g(z(C, \gamma)) \quad (1)$$

$$\text{logit}[\Pr(X = 1|C)] = \gamma^T C \quad (2)$$

where $z(C, \gamma) = \text{expit}(\gamma^T C)$ is the propensity score and $\text{expit}(a) = (1 + \exp(-a))^{-1}$. Here we write $Z = P(X = 1|C) = z(C, \gamma)$ to acknowledge that the propensity score is a known analytic function of C and the $p \times 1$ vector of regression coefficients γ .

In equation (1), the quantity β is the primary parameter of interest and models the log odds ratio for the association between Y and X given Z . The log odds ratio has well-known limitations as a measure of effect for causal inference [20], but we focus on estimating this quantity because the logit link permits flexible modelling of binary response variables. Further discussion of the problems with odds ratios in propensity score adjustment is given by Austin *et al.* [10, 11].

We let $g(Z)$ be a 5×1 vector of indicator variables, which model patient membership within one of five subclasses

$$g(Z)^T = \begin{cases} [1, 0, 0, 0, 0] & \text{if } 0 < Z < q_1 \\ [1, 1, 0, 0, 0] & \text{if } q_1 \leq Z < q_2 \\ [1, 0, 1, 0, 0] & \text{if } q_2 \leq Z < q_3 \\ [1, 0, 0, 1, 0] & \text{if } q_3 \leq Z < q_4 \\ [1, 0, 0, 0, 1] & \text{if } q_4 < Z < 1 \end{cases}$$

The quantities $\xi = (\xi_1, \dots, \xi_5)$ are corresponding regression coefficients. Thus, we model the risk of Y given X and Z as piecewise constant within the intervals $[0, q_1), [q_1, q_2), \dots, [q_4, 1]$, where the knots (q_1, q_2, q_3, q_4) are chosen *a priori*. For the EFFECT data, this model is likely to be only a rough approximation for the true mortality risk as a function of the propensity score. Nonetheless, Rosenbaum and Rubin [5] advocate stratifying on quintiles of the propensity score in order to remove 90 per cent of the bias from measured confounders, and we adopt a similar modelling strategy. BPSA can readily incorporate more flexible choices for the linear predictor $g(Z)$ through suitable modification of the MCMC algorithm. For example, we could use

cubic splines letting $\xi^T g(Z) = \xi_0 + \sum_{j=1}^3 \xi_j Z^j + \sum_{j=1}^p \xi_{j+p} (Z - q_j)_+^3$ where $(u)_+^a = u^a I(u \geq 0)$ with knots q_1, q_2, \dots, q_p , or we could let $\xi^T g(Z) = \xi_0 + \xi_1 Z$ model a simple linear predictor. To specify the knots (q_1, q_2, q_3, q_4) , we fit the logistic regression model given in equation (2) via maximum likelihood and then use the fitted values to obtain estimates of the propensity scores. The values of (q_1, q_2, q_3, q_4) are selected to define quintile groups of the estimated propensity scores. Alternatively, we could model uncertainty in the position or number of knots using Bayesian non-parametric regression techniques [13].

Equation (1) assumes that Y is conditionally independent of C given (X, Z) . This modelling assumption for control of confounding in a Bayesian analysis is formally justified by Rubin [14] and Robins and Ritov [15]. Assume that statins are prescribed in a manner such that treatment assignment is strongly ignorable given C , meaning that $p(X|Y(1), Y(0), C) = p(X|C)$ [4, 5]. This is also called the assumption of no unmeasured confounders [6]. Further, assume that $0 < Z < 1$, meaning that all patients have non-zero probability of receiving or not receiving a statin. Collectively, these conditions imply that treatment assignment is strongly ignorable given the propensity score, meaning that $p(X|Y(1), Y(0), Z) = p(X|Z)$ [4, Theorem 3]. To control confounding in a Bayesian analysis, we can calculate inferences using the modelling assumption that Y is conditionally independent of C given (X, Z) [14, 15].

2.2. Prior distributions

The quantities β , ξ , and γ are standard regression coefficients, and we assign priors in the form of normal distributions

$$\beta \sim N(0, \sigma_\beta^2)$$

$$\gamma_1, \dots, \gamma_k \sim N(0, \sigma_\gamma^2)$$

$$\xi_1, \dots, \xi_5 \sim N(0, \sigma_\xi^2)$$

where σ_β^2 , σ_γ^2 , and σ_ξ^2 are user-specified hyper-parameters. For example, we could assign $\sigma_\beta^2 = \{\log(15)/2\}^2$, to model the belief that the odds ratio for the association between X and Y given $g(Z)$ lies between 1/15 and 15 with probability 95 per cent. Alternatively, more complex modelling assumptions are possible. For example, we could supply autoregressive prior distributions for ξ_1, \dots, ξ_5 to account for possible smoothness in mortality risk between subclasses.

2.3. Posterior simulation

Let $\text{data} = \{(Y_i, X_i, C_i), i = 1, \dots, n = 4572\}$ denote the EFFECT data. We sample from the posterior density $p(\beta, \xi, \gamma | \text{data})$ using a Metropolis–Hastings algorithm, which updates parameters in blocks [13]. To motivate the approach, note that if γ is known then the propensity score $Z = z(C, \gamma) = \text{expit}(\gamma^T C)$ for each patient is also known. Posterior simulation for β and ξ may proceed via Bayesian logistic regression using the model given in equation (1). To draw from the full posterior distribution, we update successively from the conditional densities $p(\gamma | \beta, \xi, \text{data})$ and $p(\beta, \xi | \gamma, \text{data})$. Full details are given in the Appendix.

The marginal posterior distribution for β and ξ models uncertainty in the patient propensity scores. During MCMC, updating γ involves a corresponding update of Z . Patients may be grouped into different propensity score subclasses from one iteration to the next, depending on their value of

$g(Z)$. By averaging over posterior uncertainty in γ , the resulting inferences for β and ξ incorporate uncertainty in the propensity scores.

A feature of BPSA is that learning about γ is driven in part by modelling information for the outcome variable. BPSA fits both of the regression models in equations (1) and (2) simultaneously. When updating from the conditional density $p(\gamma|\beta, \xi, \text{data})$, the method assigns higher probability to the values of γ , which cluster patients with similar Y into the same propensity score subclasses. Posterior learning about mortality risk via β and ξ affects the classification of patients into subclasses. This can impact estimation as is illustrated in Sections 3 and 4.

3. BPSA ANALYSIS OF THE EFFECT DATA

We apply BPSA to the EFFECT data by first specifying values for (q_1, q_2, q_3, q_4) , which define five subclasses of the propensity score. We fit the logistic regression model in equation (2) using maximum likelihood and compute estimated propensity scores from the fitted values. We assign the quantities $q_1=0.21, q_2=0.26, q_3=0.31, q_4=0.38$, which define five quintiles of the distribution of the estimated propensity scores. We then apply BPSA to the EFFECT data. We assign $\sigma_\beta^2 = \sigma_\xi^2 = \sigma_\gamma^2 = \{\log(15)/2\}^2$ to model the prior belief that the associations between (Y, X, C) are not overly large, and we run a single MCMC chain of length 100 000 after discarding 10 000 initial iterations. Sampler convergence is assessed by simulating separate MCMC chains with overdispersed starting values and the diagnostic tools supplied in the CODA package in R [21]. To facilitate tuning of the MCMC chains in light of the different measurement scales for patient covariates, we re-scaled the laboratory values and vital signs on hospital admission to have mean zero and unit variance [13].

The analysis results are given in Table II, which contains point and 95 per cent interval estimates for the regression coefficients β, ξ , and γ . The log odds ratio for the treatment effect is equal to -0.30 with 95 per cent credible interval $(-0.50, -0.10)$ and is closer to zero compared with the crude estimate given in Section 1.1, indicating that some of the confounding between statins and mortality has been reduced. The quantities $\xi_1, \xi_2, \xi_3, \xi_4, \xi_5$ are also monotonically decreasing. This is consistent with prior information about the prescribing habits of physicians. Healthy patients are more likely to receive a statin, and consequently patients in subclass number 5 have lowest estimated mortality risk.

To put these results into perspective, we conduct an additional analysis of the EFFECT data without using Bayesian techniques. We apply a propensity score analysis (PSA), which we define as subclassification on quintiles of the estimated propensity scores. This corresponds to estimating β, ξ , and γ by fitting the regression models in equations (1) and (2) one at a time using maximum likelihood estimation. PSA is identical to BPSA, but does not acknowledge uncertainty in the propensity scores. Both methods are implemented using the same knot values of $q_1=0.21, q_2=0.26, q_3=0.31, q_4=0.38$ to define five subclasses. The results are given in the second column of Table II. BPSA and PSA give similar point estimates for β , equal to -0.30 versus -0.36 . However, the Bayesian credible interval is fully 10 per cent wider because it acknowledges uncertainty in γ . We also see differences in estimation of the nuisance parameters ξ and γ . BPSA interval estimates of $\gamma_1, \dots, \gamma_{20}$ are shorter in length compared with PSA.

To better understand the differences between BPSA and PSA, we study the covariate distributions among treated versus untreated patients within subclasses of the estimated propensity scores. For PSA, the estimated propensity score for the i th patient is given by $\hat{Z}_i^{\text{PSA}} = z(C_i, \hat{\gamma}) = \text{expit}(\hat{\gamma}^T C_i)$,

Table II. Analysis of the EFFECT data. Point and 95 per cent interval estimates for the regression coefficients β , γ , and ξ , calculated from BPSA or PSA.

Description	Parameter		BPSA		PSA
Statin effect	β	-0.30	(-0.50, -0.10)	-0.36	(-0.54, -0.18)
<i>Regression coefficient for linear predictor g{.}</i>					
y-intercept	ξ_1	0.6	(0.4, 0.9)	-0.1	(-0.3, 0.0)
Subclass number 2	ξ_2	-1.0	(-1.4, -0.7)	-0.8	(-1.0, -0.6)
Subclass number 3	ξ_3	-1.9	(-2.3, -1.6)	-1.4	(-1.7, -1.2)
Subclass number 4	ξ_4	-3.0	(-3.4, -2.6)	-1.7	(-1.9, -1.5)
Subclass number 5	ξ_5	-4.0	(-4.4, -3.6)	-2.2	(-2.5, -1.9)
<i>Demographic characteristics</i>					
Age (years)	γ_1	-0.02	(-0.03, -0.02)	-0.03	(-0.04, -0.02)
Female sex	γ_2	0.02	(-0.05, 0.09)	-0.17	(-0.33, -0.01)
<i>Presenting characteristics</i>					
Shock	γ_3	-0.37	(-1.01, 0.26)	-0.55	(-1.55, 0.45)
<i>AMI risk factors</i>					
Family history of CAD	γ_4	0.05	(-0.02, 0.12)	0.16	(0.02, 0.31)
Diabetes	γ_5	-0.10	(-0.18, -0.02)	-0.05	(-0.22, 0.12)
CVA/TIA	γ_6	-0.09	(-0.19, 0.00)	0.17	(-0.07, 0.40)
High BP	γ_7	0.09	(0.02, 0.16)	0.31	(0.17, 0.44)
Current smoker	γ_8	-0.10	(-0.19, -0.02)	-0.27	(-0.42, -0.12)
<i>Comorbidities</i>					
Angina	γ_{10}	-0.03	(-0.09, 0.04)	0.37	(0.23, 0.51)
Renal disease	γ_{11}	0.08	(-0.59, 0.76)	1.06	(0.01, 2.11)
<i>Vital signs on admission*</i>					
Systolic BP	γ_{12}	0.06	(0.02, 0.11)	0.02	(-0.08, 0.12)
Diastolic BP	γ_{13}	-0.00	(-0.05, 0.04)	-0.02	(-0.12, 0.08)
Heart rate	γ_{14}	-0.09	(-0.12, 0.06)	-0.05	(-0.12, 0.03)
Respiratory rate	γ_{15}	-0.05	(-0.08, -0.02)	-0.06	(-0.13, 0.02)
<i>Laboratory values*</i>					
White blood count	γ_{16}	-0.04	(-0.08, -0.01)	0.00	(-0.07, 0.07)
Haemoglobin	γ_{17}	0.075	(0.04, 0.11)	0.04	(-0.04, 0.12)
Sodium	γ_{18}	0.04	(0.01, 0.08)	0.05	(-0.02, 0.12)
Glucose	γ_{19}	-0.03	(-0.07, 0.01)	0.05	(-0.03, 0.12)
Creatinine	γ_{20}	-0.14	(-0.20, -0.09)	-0.07	(-0.15, 0.02)

*Continuous variables, re-scaled to have mean zero and unit variance.

where $\hat{\gamma}$ is the maximum likelihood estimate from equation (2). For BPSA, a comparable quantity is the posterior mean of $z(C_i, \gamma)$, given by

$$\hat{Z}_i^{BPSA} = E\{z(C_i, \gamma)|data\} = \int \text{expit}(\gamma^T C_i) p(\gamma|data) d\gamma$$

where $p(\gamma|data)$ denotes the marginal posterior density for γ .

Figure 1 summarizes the distribution of four important mortality risk factors, in treatment versus control, stratifying on subclasses of the estimated propensity scores. The plots on the left-hand side correspond to patients stratified on \hat{Z}_i^{PSA} , whereas the plots on the right correspond to stratification on \hat{Z}_i^{BPSA} . In Figure 1, we see that BPSA and PSA allocate patients to subclasses in different ways. Consider the variable diabetes. For PSA, the prevalence of diabetes is roughly the same in

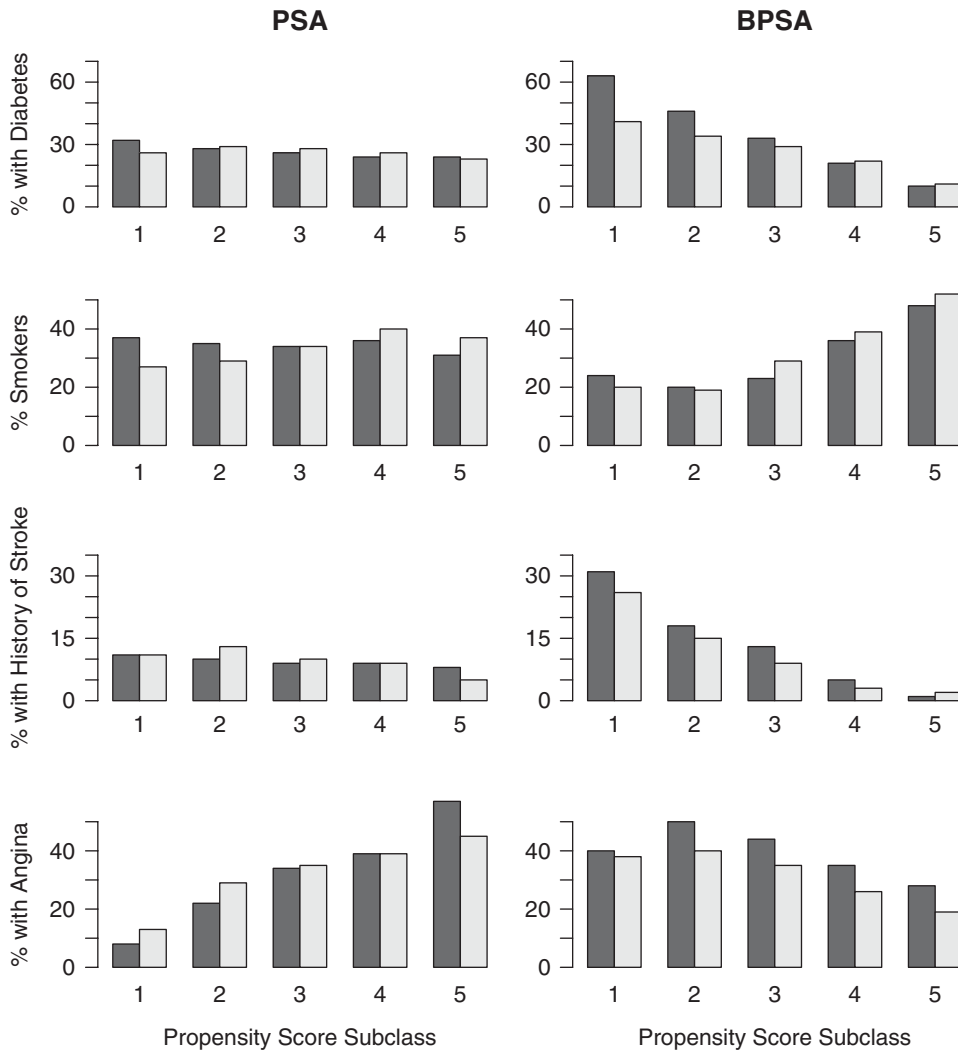


Figure 1. Prevalences of four important mortality risk factors, in treated versus untreated patients, within subclasses of the estimated propensity scores. Shaded bars refer to treated patients whereas unshaded bars refer to untreated patients.

all five subclasses. In contrast, BPSA allocates most patients with diabetes into subclass number 1. We emphasize that the intervals $[0, q_1), [q_1, q_2), \dots, [q_4, 1]$, which define the subclasses are fixed *a priori* and are the same in each analysis. PSA assigns 20 per cent of the data to each subclass. The number of subjects allocated to each of subclasses number 1 through number 5 are 914, 914, 914, 914, and 916. In contrast, for BPSA the corresponding allocations are 591, 781, 945, 1180, and 1075.

BPSA uses the outcome to inform the fit of the propensity model. In the EFFECT data, physicians prescribe statins to young healthy patients. BPSA allocates patients with mortality risk factors

such as diabetes and stroke into subclass number 1 corresponding to the lowest propensity score values lying in the interval $[0, q_1)$. During MCMC, we update γ from the conditional density $p(\gamma|\beta, \xi, \text{data})$. Learning about ξ and β affects estimation of γ . Conceptually, BPSA optimizes the likelihood function for the response multiplied by the likelihood function for treatment. Fitting the propensity model comes at the expense of the outcome model. Using the outcome variable in order to inform the fit of the propensity model is an unusual approach to estimation, and we review the different perspectives in the discussion in Section 6.

The results in Table II motivate questions about the estimation strategies used by BPSA and PSA. Both methods make exactly the same modelling assumptions, but they handle information in different ways. There may be benefits and trade-offs to BPSA. If equation (1) is sufficiently non-parametric and accurately models the relationship between mortality and the propensity score, then BPSA may give improved estimation because it makes fuller use of modelling assumptions. Conversely, PSA may be more robust because it estimates the propensity scores using only the marginal model for $\Pr(X = 1|C)$. We explore these issues in Section 4 using simulations.

4. SIMULATIONS STUDIES OF THE PERFORMANCE OF BPSA AND PSA

The results from the EFFECT data example show that modelling uncertainty in the propensity score can increase the length of interval estimates for the treatment effect. However, the question remains whether or not this is a general phenomenon in data sets with confounding. In what settings can we expect that there will be large uncertainty, and does ignoring it harm estimation?

We study the performance of interval estimates using simulations. In Section 4.1, we apply BPSA to synthetic data where the log odds of Y given X and C follows equation (1) and is constant within subclasses of the propensity score. A potential weakness of BPSA is that it incorporates modelling assumptions about the relationship between Y and Z when estimating the propensity scores. This raises the effect of the impact of model misspecification. To study the impact of this estimation approach, Section 4.2 investigates the performance of BPSA when applied to synthetic data where the log odds of Y depends linearly on X and C . Such models have been used in the past to evaluate the performance of PSA in simulations with binary outcomes [9–12].

4.1. Simulation study when the log odds of Y given X and C follows equation (1)

4.1.1. Simulation design. We consider the case where C has four continuous components. We simulate ensembles of 400 data sets of sample size $n = 1000$, for three different choices of model parameters, denoted as Designs A, B or C. Data are generated using the following algorithm:

1. Generate C_1, \dots, C_n where the first component is equal to one and the latter four components are independently drawn as $N(0, 1)$ random variables.
2. For fixed γ , generate X_1, \dots, X_n using the logistic regression model of equation (2).
3. Given $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$, we have $\gamma^T C \sim N(\gamma_0, \gamma_1^2 + \gamma_2^2 + \gamma_3^2 + \gamma_4^2)$. The values (q_1, q_2, q_3, q_4) defining the true quintiles of the propensity score are $q_k = \text{expit}\{\gamma_0 + \sqrt{\gamma_1^2 + \gamma_2^2 + \gamma_3^2 + \gamma_4^2} \times \Phi^{-1}(0.2k)\}$ for $k = 1, 2, 3, 4$, where $\Phi^{-1}(\cdot)$ is the quantile function of a $N(0, 1)$ random variable. Generate Y_1, \dots, Y_n using the logistic regression model in equation (1), for fixed (β, ξ) .

Table III. Catalogue of simulation designs. Each row corresponds to the true parameter values that are used to generate synthetic data for the given simulation design.

Design	β	ξ	γ
<i>Simulations for Section 4.1</i>			
A	0	(-4, -2, 0, 2, 4)	(0.1, 0.1, 0.1, 0.1, 0.1)
B	0	(-2, -1, 0, 1, 2)	(0.1, 0.1, 0.1, 0.1, 0.1)
C	0	(-2, -1, 0, 1, 2)	(0.5, 0.5, 0.5, 0.5, 0.5)
<i>Simulations for Section 4.2</i>			
D	0	(2, 2, 2, 2, 2)	(0.1, 0.1, 0.1, 0.1, 0.1)
E	0	(1, 1, 1, 1, 1)	(0.1, 0.1, 0.1, 0.1, 0.1)
F	0	(1, 1, 1, 1, 1)	(0.5, 0.5, 0.5, 0.5, 0.5)

We analyze the data sets using PSA and BPSA. Sampler convergence is assessed using separate trial MCMC runs.

Table III gives a catalogue of the fixed values of β , ξ , and γ that are used to generate the synthetic data. Designs A and B model the setting where there are weak associations between X and C , as is typical in the EFFECT data set (see Table II). Design C models stronger associations. We set $\beta=0$ to model no causal effect of X on Y within levels of the propensity score. The choice of true fixed values for ξ is guided by the results in Table II and models a monotonically varying association between Y and $g(Z)$.

4.1.2. Results. Table IV summarizes the performance of point and 80 per cent interval estimates for β , ξ , and γ calculated from BPSA or PSA applied to ensembles of 400 synthetic data sets, which are generated under Designs A, B or C. Beneath the heading of point estimation, the first two columns give information about bias and contain the difference between the sample means of the ensembles of the point estimates and the true parameter value. The third and fourth columns contain the estimated relative efficiency and mean squared error (MSE). These quantities are calculated as

$$\text{Estimated relative efficiency} = \frac{\text{Sample variance of BPSA point estimates}}{\text{Sample variance of PSA point estimates}}$$

and

$$\text{Estimated relative MSE} = \frac{\text{Sample average squared error of BPSA point estimates}}{\text{Sample average squared error of PSA point estimates}}$$

To aid with interpretation of results, we calculated *simulation standard errors* for the relative efficiency and relative MSE estimates via the bootstrap. In Table IV, quantities denoted with a '‡' imply that a 90 per cent bootstrapped confidence interval for the true relative efficiency or MSE excludes 1. The final four columns contain empirical coverage probabilities and average lengths of 80 per cent interval estimates.

Table IV illustrates that BPSA 80 per cent interval estimates for the treatment effect β are wider compared with PSA when the association between treatment and confounders is weak. In Design A, the average length is 0.59 for BPSA compared with 0.54 for PSA, corresponding to a 10 per cent increase. The effect of modelling uncertainty in the estimated propensity scores is also reflected in point estimation of β . For Design A, the relative efficiency is 1.25 implying that BPSA has a 25 per cent larger variance. The difference in efficiency is less pronounced when the

Table IV. Performance of point and 80 per cent interval estimates calculated from BPSA and PSA applied to ensembles of 400 synthetic data sets generated under Designs A, B or C.

Fixed parameter	Point estimation				80 per cent interval estimates			
	Bias		Relative*		Coverage [†]		Length	
	BPSA	PSA	Efficiency	MSE	BPSA (per cent)	PSA (per cent)	BPSA	PSA
Design A, $\xi = (-4, -2, 0, 2, 4)$								
$\beta = 0$	0.02	0.05	1.25 [‡]	1.20 [‡]	78	78	0.59	0.54
$\gamma_0 = 0.1$	-0.03	0.00	0.55 [‡]	0.77 [‡]	64	78	0.05	0.16
$\gamma_1 = 0.1$	0.01	0.00	0.20 [‡]	0.20 [‡]	69	78	0.04	0.16
$\gamma_2 = 0.1$	0.01	0.00	0.23 [‡]	0.23 [‡]	72	84	0.03	0.16
$\gamma_3 = 0.1$	0.01	0.00	0.19 [‡]	0.20 [‡]	72	81	0.03	0.16
$\gamma_4 = 0.1$	0.01	-0.01	0.18 [‡]	0.18 [‡]	72	76	0.03	0.16
Design B, $\xi = (-2, -1, 0, 1, 2)$								
$\beta = 0$	0.01	0.03	1.14 [‡]	1.11 [‡]	78	77	0.44	0.42
$\gamma_0 = 0.1$	-0.01	0.00	0.52 [‡]	0.57 [‡]	72	80	0.07	0.16
$\gamma_1 = 0.1$	0.01	0.00	0.28 [‡]	0.32 [‡]	70	79	0.05	0.16
$\gamma_2 = 0.1$	0.01	0.00	0.30 [‡]	0.33 [‡]	73	81	0.05	0.16
$\gamma_3 = 0.1$	0.01	0.00	0.35 [‡]	0.38 [‡]	73	82	0.05	0.16
$\gamma_4 = 0.1$	0.01	0.00	0.27 [‡]	0.29 [‡]	72	80	0.05	0.16
Design C, $\xi = (-2, -1, 0, 1, 2)$								
$\beta = 0$	-0.01	-0.01	1.01	1.01	80	79	0.52	0.51
$\gamma_0 = 0.5$	0.00	0.00	0.52 [‡]	0.98	77	80	0.11	0.19
$\gamma_1 = 0.5$	0.00	0.01	0.36 [‡]	0.96	78	80	0.11	0.19
$\gamma_2 = 0.5$	0.01	0.00	0.41 [‡]	1.00	79	81	0.11	0.19
$\gamma_3 = 0.5$	0.01	0.00	0.35 [‡]	1.00	74	77	0.11	0.19
$\gamma_4 = 0.5$	0.00	0.00	0.40 [‡]	0.99	79	80	0.11	0.19

*Ratio of BPSA to PSA.

[†]Simulation standard errors for coverage estimates are less than $\sqrt{0.5(1-0.5)}/400=2.5$ per cent.

[‡]Quantity differs from 1, p -value <0.1 .

true fixed values of $\gamma = (\gamma_0, \dots, \gamma_4)$ are large. For Designs B and C, the relative efficiencies for point estimates of β are only 1.14 and 1.01, respectively. Intuitively, a weak association between X and C makes it difficult for PSA to discriminate between patients based on their propensity for treatment. There is greater uncertainty in classifying patients into subclasses.

Remarkably, ignoring uncertainty in the estimated propensity scores appears to have no adverse effect on coverage probability. PSA interval estimates are shorter on average compared with BPSA, and yet they retain nominal 80 per cent coverage levels. PSA point estimates of β are more efficient and have smaller MSE. Contrary to intuition, substituting estimates in place of the true propensity scores does not yield inferences, which are falsely precise.

To better understand the performance of BPSA, we can study estimates of the nuisance parameter γ , which models the propensity scores. Design A of Table IV shows that BPSA estimates of γ are slightly biased, with coverage probability near 70 per cent and well below the nominal level.

Because the BPSA model is identical to the true data generating process in the simulation, we know that in large samples this bias must vanish. For Bayesian inference in general, the posterior mean is asymptotically consistent to the parameter value when the model is specified correctly and under standard regularity conditions [13].

Table IV shows that Bayesian point estimates of γ are more efficient compared with PSA, with relative efficiencies generally below 0.50. As a result, under Designs A and B we see that BPSA point estimates of γ have lower MSE, despite the small bias. The intuitive explanation is that when the distribution of Y is a heterogeneous function of Z , this means that the outcome variable carries information about the propensity score. BPSA uses this information and the resulting point estimates of γ are more efficient compared with PSA. These findings are echoed in the EFFECT data analysis of Section 3. In Table II, BPSA interval estimates for $\gamma_1, \dots, \gamma_{20}$ are narrower than PSA. Section 6 reviews the different viewpoints on the role of the outcome in propensity score modelling.

Because the true value of γ is known by design, simulations allow us to study the characteristics of PSA using the true propensity scores. Accordingly, Table V summarizes the performance of PSA estimates for β in the case when γ is known. Table V confirms that using true rather than estimated propensity scores harms the efficiency of PSA point estimates for the treatment effect β [2]. For all simulation designs, the relative efficiencies are greater than one. Furthermore, a comparison of Tables IV and V shows that the performance of BPSA interval estimates for β is not worse than PSA using the true propensity scores. The intervals have the same average length and both have nominal coverage levels.

4.2. Simulation study when the log odds of Y given X and C depends linearly on X and C

4.2.1. *Simulation design.* To explore sensitivity to modelling assumptions for the outcome variable, we repeat the simulations using exactly the same algorithm as in Section 4.1.1, except that in

Table V. Performance of point and 80 per cent interval estimates of the treatment effect β calculated from PSA using the true propensity scores and applied to ensembles of 400 synthetic data sets generated under Designs A through F.

Design	True parameter	Point estimation			80 per cent interval estimates	
		Bias	Relative*		Coverage [†] (per cent)	Length
			Efficiency	MSE		
Design A	$\beta=0$	0.00	1.16 [‡]	1.10 [‡]	80	0.59
Design B	$\beta=0$	0.00	1.03	1.01	80	0.44
Design C	$\beta=0$	0.00	1.02	1.02	79	0.52
Design D	$\beta=0$	0.03	2.22 [‡]	2.17 [‡]	80	0.55
Design E	$\beta=0$	0.02	1.45 [‡]	1.42 [‡]	78	0.43
Design F	$\beta=0$	0.09	1.03	0.95	74	0.45

*PSA using true propensity scores relative to PSA using estimated propensity scores.

[†]Simulation standard errors for coverage estimates are less than $\sqrt{0.5(1-0.5)}/400=2.5$ per cent.

[‡]Quantity differs from 1, p -value<0.1.

Table VI. Performance of point and 80 per cent interval estimates calculated from BPSA or PSA applied to ensembles of 400 synthetic data sets generated under Designs D, E or F.

True parameter	Point estimation				80 per cent interval estimates			
	Bias		Relative*		Coverage [†]		Length	
	BPSA	PSA	Efficiency	MSE	BPSA (per cent)	PSA (per cent)	BPSA	PSA
Design D, $\xi = (2, \dots, 2)$								
$\beta = 0$	0.00	0.00	1.78 [‡]	1.69 [‡]	82	82	0.57	0.46
$\gamma_0 = 0.1$	0.06	0.00	0.66 [‡]	1.67 [‡]	7	79	0.03	0.16
$\gamma_1 = 0.1$	0.03	0.00	0.27 [‡]	0.44 [‡]	18	78	0.03	0.16
$\gamma_2 = 0.1$	0.03	0.00	0.26 [‡]	0.43 [‡]	20	79	0.03	0.16
$\gamma_3 = 0.1$	0.03	0.00	0.28 [‡]	0.46 [‡]	20	82	0.03	0.16
$\gamma_4 = 0.1$	0.03	0.00	0.30 [‡]	0.48 [‡]	21	81	0.03	0.16
Design E, $\xi = (1, \dots, 1)$								
$\beta = 0$	0.01	0.03	1.41 [‡]	1.36 [‡]	80	83	0.44	0.40
$\gamma_0 = 0.1$	0.03	0.00	0.46 [‡]	0.69 [‡]	21	79	0.03	0.16
$\gamma_1 = 0.1$	0.00	0.00	0.21 [‡]	0.21 [‡]	46	78	0.03	0.16
$\gamma_2 = 0.1$	0.00	0.00	0.20 [‡]	0.21 [‡]	44	81	0.03	0.16
$\gamma_3 = 0.1$	0.00	0.00	0.21 [‡]	0.21 [‡]	44	80	0.03	0.16
$\gamma_4 = 0.1$	0.00	0.00	0.21 [‡]	0.21 [‡]	48	80	0.03	0.16
Design F, $\xi = (1, \dots, 1)$								
$\beta = 0$	0.08	0.10	1.00	0.89 [‡]	77	73	0.46	0.45
$\gamma_0 = 0.5$	0.04	0.00	1.81 [‡]	1.24 [‡]	47	81	0.12	0.19
$\gamma_1 = 0.5$	-0.01	0.01	0.71 [‡]	0.92 [‡]	55	81	0.11	0.19
$\gamma_2 = 0.5$	-0.01	0.01	0.77 [‡]	0.90 [‡]	51	81	0.11	0.19
$\gamma_3 = 0.5$	-0.01	0.00	0.68 [‡]	0.95	58	79	0.11	0.19
$\gamma_4 = 0.5$	-0.01	0.01	0.63 [‡]	0.90 [‡]	58	82	0.11	0.19

* Ratio of BPSA to PSA.

[†] Simulation standard errors for coverage estimates are less than $\sqrt{0.5(1-0.5)}/400 = 2.5$ per cent.

[‡]Quantity differs from 1, p -value < 0.1.

Step 3 we generate Y_1, \dots, Y_n using the logistic regression model

$$\text{logit}[\Pr(Y = 1|X, C)] = \beta X + \xi^T C$$

Similar models are used by Drake, Austin *et al.* and Brookhart *et al.* [9–12] to evaluate the performance of PSA in simulation studies. We analyze the data sets with PSA and BPSA using the values of (q_1, q_2, q_3, q_4) , which define the true quintiles of the propensity score. The parameter values used in this simulation correspond to Designs D, E, and F in Table III. Again, we consider settings with weak or strong association between X , Y , and C . We set β equal to zero to model no association between X and Y given C .

4.2.2. *Results.* Table VI presents the simulation results and shows that the basic features of BPSA are insensitive to misspecification of the model for the outcome. When the true value of γ is small, modelling uncertainty in the propensity scores has a large impact on treatment effect estimation. BPSA interval estimates are substantially wider compared with PSA, but show no noticeable improvement in coverage probability. BPSA estimates of the nuisance parameter γ are biased compared with PSA, but generally more efficient as a result of information flow from the outcome variable. The trade-off between bias and efficiency leads to an overall reduction in MSE. Thus, while the subclassification model in equation (1) is incorrect in this simulation, it sufficiently approximates the true data generating process such that the performance characteristics BPSA and PSA are unaffected.

5. PREDICTIVE PERFORMANCE IN THE EFFECT DATA

In the EFFECT data, the relationship between mortality and the propensity score is unknown and equation (1) is only a rough approximation. It is unclear whether we should use the full Bayesian approach to fit the data, or instead adopt a more conservative estimation strategy by calculating the propensity scores from the marginal model for treatment. If the outcome model is wrong, then the Bayesian approach might adversely impact estimation. One way to explore model fitting in the EFFECT data is through prediction error. We can build predictive models for mortality using BPSA or PSA and then assess the quality of the predictions using cross-validation.

The EFFECT data include 4572 patients discharged from hospital between 1999 and 2000. However, data from patients discharged the following year are also available to yield a total sample size of 9171. We investigate prediction error for mortality using cross-validation by randomly splitting the entire data set in half. Denote a collection of data of size $n = 4500$ from the population under study as $\text{data} = \{(Y_i, X_i, C_i), i = 1, \dots, n, \}$. Let (Y^*, X^*, C^*) denote the data for a future patient from the same population for whom only X^* and C^* are observed. Let $\hat{\beta}$, $\hat{\xi}$, and $\hat{\gamma}$ denote the point estimates for model parameters from PSA applied to data, and define

$$\hat{Y}^{\text{PSA}} = \text{expit}\{\hat{\beta}X^* + \hat{\xi}^T g(z(C^*, \hat{\gamma}))\}$$

as the predictive model from PSA, which estimates $P(Y^* = 1|X^*, C^*)$. Define

$$\hat{Y}^{\text{BPSA}} = \int \int \int \text{expit}\{\beta X^* + \xi^T g(z(C^*, \gamma))\} p(\beta, \xi, \gamma | \text{data}) d\beta d\xi d\gamma$$

as the predictive model from BPSA for estimation of $P(Y^* = 1|X^*, C^*)$, where $p(\beta, \xi, \gamma | \text{data})$ is the posterior distribution. The quantity \hat{Y}^{PSA} is a prediction based on substitution of $(\hat{\beta}, \hat{\xi}, \hat{\gamma})$ into equation (1), while \hat{Y}^{BPSA} is the posterior mean from the predictive distribution for Y^* . The estimate \hat{Y}^{BPSA} acknowledges uncertainty in the subclass of the patient, while \hat{Y}^{PSA} does not.

We quantify prediction error using the loss function

$$L(\hat{Y}, Y^*) = -[Y^* \log(\hat{Y}) + (1 - Y^*) \log(1 - \hat{Y})]$$

which is the logarithm of the observed ordinate of the predictive model for Y^* [13]. The prediction error is then $E\{L(\hat{Y}, Y^*)\}$. This quantity is closely connected to the model deviance, which is a

Table VII. Predictive error for mortality in the EFFECT data. The quantities $\hat{\phi}_{\text{BPSA}}$ and $\hat{\phi}_{\text{PSA}}$ are calculated from five different random splittings into build data ($n=4500$) and test data ($n=4671$).

Random splitting	Prediction error	
	$\hat{\phi}_{\text{BPSA}}$	$\hat{\phi}_{\text{PSA}}$
1	0.40	0.47
2	0.38	0.45
3	0.39	0.46
4	0.39	0.45
5	0.39	0.46
Average error	0.39	0.46

standard measure of goodness of fit for dichotomous regression [13]. Predictions with small error have the appealing property that they assign high probability to the data that are actually observed.

To estimate the prediction errors $E\{L(\hat{Y}^{\text{BPSA}}, Y^*)\}$ and $E\{L(\hat{Y}^{\text{PSA}}, Y^*)\}$ for the EFFECT data, we randomly select a data set of size $n=4500$ from the total sample of 9171. We analyze the data to obtain predictive models and then evaluate these predictions in the remaining 4671 patients by calculating the estimates

$$\hat{\phi}_{\text{BPSA}} = -\frac{1}{4671} \sum_{i=1}^{4671} [Y_i^* \log(\hat{Y}_i^{\text{BPSA}}) + (1 - Y_i^*) \log(1 - \hat{Y}_i^{\text{BPSA}})]$$

and

$$\hat{\phi}_{\text{PSA}} = -\frac{1}{4671} \sum_{i=1}^{4671} [Y_i^* \log(\hat{Y}_i^{\text{PSA}}) + (1 - Y_i^*) \log(1 - \hat{Y}_i^{\text{PSA}})]$$

where the summation i is over observations in the remaining data. The quantities $\hat{\phi}_{\text{BPSA}}$ and $\hat{\phi}_{\text{PSA}}$ are unbiased estimates of $E\{L(\hat{Y}^{\text{BPSA}}, Y^*) | \text{data}\}$ and $E\{L(\hat{Y}^{\text{PSA}}, Y^*) | \text{data}\}$, respectively. To fully characterize $E\{L(\hat{Y}^{\text{BPSA}}, Y^*)\}$ we can repeatedly split the data and examine the sequence of $\hat{\phi}_{\text{BPSA}}$ and $\hat{\phi}_{\text{PSA}}$. Table VII presents the prediction error estimates $\hat{\phi}_{\text{BPSA}}$ and $\hat{\phi}_{\text{PSA}}$ for five random splittings of the data. Each row gives $\hat{\phi}_{\text{BPSA}}$ and $\hat{\phi}_{\text{PSA}}$ for one random splitting. Standard errors for the estimates are calculated as the sample standard deviation of the replicates of $L(\hat{Y}, Y^*)$, divided by $\sqrt{4671}$, and they are less than 0.008. For each row in the table, $\hat{\phi}_{\text{BPSA}}$ is smaller than $\hat{\phi}_{\text{PSA}}$. Thus, the prediction estimates \hat{Y}^{BPSA} have smaller error compared with \hat{Y}^{PSA} . The decrease in error is equal to $0.46 - 0.39 = 0.07$ on average, meaning that BPSA gives an $\exp(0.07) = 7$ per cent improvement in correct prediction of patient mortality.

This gives some reassurance that the manner in which BPSA estimates propensity scores yields a satisfactory fit for the outcome regression model. When estimating the propensity scores at the same time as other parameters, BPSA gains flexibility in the way that patients are grouped into propensity score subclasses. Mortality risk within treatment groups is strongly dependent on the propensity score, and BPSA uses this information in order to group patients based on health status.

While the resulting propensity score estimates differ from those of PSA, the method gives an improved fit for the data with smaller prediction error for mortality.

6. DISCUSSION

In this paper we present an investigation of modelling uncertainty in the propensity scores using Bayesian techniques. We focus on the specific case of subclassification on quintiles of the propensity score in observational studies with a dichotomous treatment, dichotomous outcome, and measured confounders where the log odds ratio is the measure of effect. Using simulations, we demonstrate that when the association between the treatment and confounders is weak, as is the case in the EFFECT data example, then there may be large uncertainty in the estimated propensity scores. In a Bayesian analysis, the interval estimate for the treatment effect is 10 per cent wider. Remarkably, simulations show that ignoring uncertainty in the estimates propensity scores has no adverse impact on coverage probability. PSA interval estimates for the treatment effect are shorter on average compared with BPSA, and yet they retain nominal coverage levels. PSA point estimates for the treatment effect are also more efficient with smaller MSE. There is a large literature on the merits of using estimated rather than true propensity scores for control of confounding (see [2] for review). Propensity scores that have been estimated from the marginal model $\Pr(X = 1|C)$ have the property that they allow adjustment for chance imbalances between treatment groups due to random variation, thereby improving the efficiency of treatment effect estimates [2].

A feature of BPSA is that it uses the outcome variable to inform the fit of the propensity model. In regression adjustment for the propensity score, the decision to stratify on subclasses may be driven by prior beliefs about the relationship between Y and Z given X . For the EFFECT data, patients with a high propensity scores are more healthy. BPSA makes formal use of this modelling information when estimating the propensity scores.

Using the outcome variable to fit the propensity model may give more efficient estimation of the propensity scores. While these are merely nuisance parameters, the findings are interesting nonetheless. The simulations of Section 4 show that when the outcome depends heavily on the propensity score, it carries information that can assist estimation. Even if the subclassification model is misspecified, BPSA performs well with respect to estimation of γ , provided that the model approximates the relationship between the outcome and propensity score. Additionally, Section 5 studies prediction error in the EFFECT data and confirms that BPSA propensity scores estimates yield a classification of subjects into quintiles, which gives a better overall fit for the data. Simultaneous model fitting for treatment and outcome gives added flexibility in estimation of γ . Similar findings have been reported in other Bayesian latent variable modelling applications, including measurement error [22] and Bayesian modelling averaging [23]. Acknowledging uncertainty in nuisance parameter estimates tends to improve predictive performance.

There are different viewpoints in the literature on using the outcome to assist estimation of the propensity scores. Rubin argues that propensity score modelling should be conducted without access to the outcome data [2]. This perspective seeks to replicate a randomized trial by producing treatment groups that are similar with respect to measured confounders and without consideration for the outcome under study. Other authors emphasize the value of prior information about the strength of the association between confounders and the outcome in propensity score model building [10–12]. If a covariate is not an outcome risk factor, then it may not matter if it is a powerful predictor of treatment assignment, and it can be safely ignored in the propensity model

without introducing confounding bias. The outcome variable may contain valuable information for estimating the propensity scores.

APPENDIX: MARKOV CHAIN MONTE CARLO

We outline a method for sampling from the posterior density $p(\beta, \xi, \gamma|\text{data})$ using the Metropolis–Hastings algorithm to update successively from $p(\gamma|\beta, \xi, \text{data})$ and $p(\beta, \xi|\gamma, \text{data})$. We have $p(\beta, \xi, \gamma, \text{data}) \propto \prod_{i=1}^n p(Y_i|X_i, C_i, \beta, \xi, \gamma)p(X_i|C_i, \gamma)p(\beta)p(\xi)p(\gamma)$. The conditional density $p(\gamma|\beta, \xi, \text{data})$ obeys the proportionality

$$\begin{aligned} p(\gamma|\beta, \xi, \text{data}) &\propto \prod_{i=1}^n p(Y_i|X_i, C_i, \beta, \xi, \gamma)p(X_i|C_i, \gamma)p(\gamma) \\ &= \prod_{i=1}^n \left[\frac{\exp\{Y_i(\beta X_i + \xi^T g(z(C_i, \gamma)))\}}{1 + \exp\{\beta X_i + \xi^T g(z(C_i, \gamma))\}} \times \frac{\exp\{X_i(\gamma^T C_i)\}}{1 + \exp\{\gamma^T C_i\}} \right] \times p(\gamma) \end{aligned}$$

This density is not proportional to $\prod_{i=1}^n p(X_i|C_i, \gamma)p(\gamma)$. Therefore, updating γ does not consist of sampling from the posterior distribution of the regression coefficients from logistic regression of X on C . The density $p(\gamma|\beta, \xi, \text{data})$ assigns high probability to values of γ for which the quantities $g(z(C_1, \gamma)), \dots, g(z(C_n, \gamma))$ yield the best fit for the regression model for the outcome. To update γ , we use a proposal distribution based on the approximation

$$p(\gamma|\beta, \xi, \text{data}) \approx \prod_{i=1}^n p(X_i|C_i, \gamma)p(\gamma)$$

The proposal is equal to the current value of γ plus random noise given by a draw from a multivariate normal distribution of dimension p with mean zero and covariance matrix equal to the inverse of the observed information from logistic regression of X on C . This approach permits simultaneous updating of all components of γ and is empirically found to give satisfactory acceptance rates.

To update β and ξ given γ and data, we have

$$\begin{aligned} p(\beta, \xi|\gamma, \text{data}) &\propto \prod_{i=1}^n p(Y_i|X_i, C_i, \beta, \xi, \gamma)p(\beta)p(\xi) \\ &= \prod_{i=1}^n \frac{\exp\{Y_i(\beta X_i + \xi^T g(z(C_i, \gamma)))\}}{1 + \exp\{\beta X_i + \xi^T g(z(C_i, \gamma))\}} \times p(\beta)p(\xi) \end{aligned}$$

This is the posterior distribution for Bayesian logistic regression of Y on X and $g(Z)$, and posterior simulation is accomplished using standard techniques [13].

ACKNOWLEDGEMENTS

The EFFECT study is funded by a Canadian Institutes of Health Research Team Grant in Cardiovascular Outcomes Research. Additional funding for this project came from an operating grant (Grant No. NA 5703) from the Heart and Stroke Foundation of Ontario. Lawrence McCandless is supported by a training award from the British Columbia Michael Smith Foundation for Health Research.

REFERENCES

1. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine* 2005; **25**:2084–2106.
2. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine* 2007; **26**:20–36.
3. Stümer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ. Analytic strategies to adjust for confounding using exposure propensity scores and disease risk scores: nonsteroidal anti-inflammatory drugs and short-term mortality in the elderly. *American Journal of Epidemiology* 2005; **161**:891–899.
4. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–57.
5. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
6. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 2004; **23**:2937–2960.
7. Tu W, Zhou XH. A bootstrap confidence interval procedure for the treatment effect using propensity score subclassification. *Health Services and Outcomes Research Methodology* 2002; **3**:135–147.
8. Hirano K, Imbens GW. The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives*, Gelman A, Meng X (eds). Wiley: New York, 2004; 71–84.
9. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 1993; **49**:1231–1236.
10. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine* 2006; **26**:3078–3094.
11. Austin PC, Grootendorst P, Normand ST, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Statistics in Medicine* 2006; **26**:754–768.
12. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stümer T. Variable selection for propensity score models. *American Journal of Epidemiology* 2006; **163**:1149.
13. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis* (2nd edn). Chapman Hall/CRC: New York, Boca Raton, 2004.
14. Rubin DB. The use of propensity scores in applied Bayesian inference. In *Bayesian Statistics 2*, Bernardo JM, De Groot MH, Lindley DV, Smith AFM (eds). Valencia University Press, North-Holland: Amsterdam, 1985; 63–72.
15. Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* 1997; **16**:285–318.
16. Little RJA. To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association* 2004; **99**:546–556.
17. Ko D, Mamdani M, Alter D. Lipid-lowering therapy with statins in high-risk elderly patients: the treatment-risk paradox. *Journal of the American Medical Association* 2004; **291**:1864–1870.
18. Glynn R, Schneeweiss S, Wang P, Levin R, Avorn J. Selective prescribing led to overestimation of the benefits of lipid-lowering drugs. *Journal of Clinical Epidemiology* 2002; **59**:819–828.
19. Stenestrand U, Wallentin L. Early statin treatment following acute myocardial infarction and 1-year survival. *Journal of the American Medical Association* 2001; **285**:430–436.
20. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology* 1987; **125**:761–768.
21. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, 2004. ISBN 3-900051-00-3. URL <http://www.R-project.org>.
22. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman Hall/CRC: New York, Boca Raton, 2003.
23. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Statistical Science* 1999; **14**:382–417.