A Bayesian view of doubly robust causal inference

O. Saarela*, L. R. Belzile† AND D. A. Stephens‡

Abstract

In causal inference confounding may be controlled either through regression adjustment in an outcome model, or through propensity score adjustment or inverse probability of treatment weighting, or both. The latter approaches, which are based on modelling of the treatment assignment mechanism and their doubly robust extensions have been difficult to motivate using formal Bayesian arguments; in principle, for likelihood-based inferences, the treatment assignment model can play no part in inferences concerning the expected outcomes if the models are assumed to be correctly specified. On the other hand, forcing dependency between the outcome and treatment assignment models by allowing the former to be misspecified results in loss of the balancing property of the propensity scores and the loss of any double robustness. In this paper, we explain in the framework of misspecified models why doubly robust inferences cannot arise from purely likelihood-based arguments, and demonstrate this through simulations. As an alternative to Bayesian propensity score analysis, we propose a Bayesian posterior predictive approach for constructing doubly robust estimation procedures. Our approach appropriately decouples the outcome and treatment assignment models by incorporating the inverse treatment assignment probabilities in Bayesian causal inferences as importance sampling weights in Monte Carlo integration.

A revised version of this article has been accepted for publication in *Biometrika*, published by Oxford University Press.
Saarela, O., Belzile, L. R. and D. A. Stephens. A Bayesian view of doubly robust causal inference, *Biometrika* (2016), 103 (3): 667-681. doi:10.1093/biomet/asw025.

1. INTRODUCTION

In causal inference contexts, confounding is most often controlled by one of two approaches: first, by conditioning on the confounders in a regression model for the outcome in an *outcome regression* model; secondly, by modelling the treatment assignment mechanism to obtain so-called *propensity score* values, and then using these scores to construct strata within the observed sample, or a pseudo-sample from a hypothetical population, within which treatment assignment is not confounded. This pseudo-sample can be obtained via inverse probability of treatment weighting of the original sample, analogously to survey sampling procedures. The outcome regression adjustment method requires correct specification of the regression function in order to obtain consistent inference; this may be achieved in practice using flexible regression strategies and complex functions of typically a large number of covariates. The propensity score adjustment methods focus principally on the specification of the treatment assignment model, which may be similarly flexible or complex. Under either approach, sufficient control of confounding is therefore dependent on possibly unverifiable modelling assumptions. This has motivated the development of *doubly robust* methods in which both model components are specified, but only one of them needs to be correctly specified to sufficiently control for confounding.

Adjustments that depend on the propensity score using regression or reweighting are not easy to interpret from the Bayesian perspective, since Bayesian inferences are naturally based on modelling of the outcome, with modelling of the treatment assignment playing no role in inference relating to the outcome/treatment relationship (Robins and Ritov, 1997). Gustafson (2012) attempted a Bayesian interpretation as a compromise between a saturated outcome model and a parametric one; however, the treatment assignment model did not feature in this interpretation. Scharfstein et al. (1999) and Bang and

*Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, Ontario M5T 3M7, Canada. olli.saarela@utoronto.ca
†École Polytechnique Fédérale de Lausanne, EPFL-SB-MATHAA-STAT, Station 8, CH-1015 Lausanne, Switzerland. leo.belzile@epfl.ch
‡Department of Mathematics and Statistics, McGill University, Montreal, Quebec H3A 2K6, Canada. d.stephens@math.mcgill.ca

Robins (2005) pointed out a connection between a doubly robust estimator and a model-based estimator; the two are equivalent if the outcome model in the latter features the so called clever covariate, a particular function of the propensity score.

Separately from the above developments, there has been a body of work studying Bayesian versions of propensity score adjustment to control for confounding (Hoshino, 2008; McCandless et al., 2009a,b, 2010, 2012; An, 2010; Kaplan and Chen, 2012; Zigler et al., 2013; Chen and Kaplan, 2015; Graham et al., 2015). These approaches require one of two kinds of compromises; they either force a parametrization which makes the outcome and treatment assignment models dependent, thus losing the balancing property of the propensity score, or cut such a feedback in which case the inference procedures are no longer Bayesian. Because of these difficulties, some authors (e.g. Achy-Brou et al., 2010, p. 828) have been content to fix the propensity scores to their best estimates in model-based inferences, without attempting to jointly estimate the two model components. In an alternative approach, Wang et al. (2012) and Zigler and Dominici (2014) have suggested connecting the outcome and treatment assignment models through the prior distribution in order to incorporate the uncertainty in confounder selection. Herein we do not consider model uncertainty, but rather, concentrate on inferences with a priori specified outcome and treatment assignment models.

The purpose of this paper is to clarify the theoretical and practical motivations for Bayesian propensity score adjustment, and the relationships between the different methods proposed for this, which have not been fully explored previously. We address these in the context of double adjustment for both the potential confounders and the propensity score, and argue that the problem cannot be properly understood without considering it in the framework of misspecified models. To provide an alternative to Bayesian propensity score adjustment, we propose deriving Bayesian versions of various inverse probability of treatment weighted estimators, including inverse probability of treatment weighted outcome regression and the semi-parametric double robust estimator, through posterior predictive expectations, with the weights introduced as importance sampling weights in Monte Carlo integration.

2. PRELIMINARIES

2. *Notation and assumptions*

For simplicity, we consider the case of a single binary treatment, and defer discussion of the longitudinal, multiple exposure and continuous cases to the Discussion. Let the random vectors X_i represent a set of pre-treatment covariate measurements, Z_i a binary treatment allocation indicator, and Y_i an outcome for individual i, measured after sufficient time has passed since administering the treatment. We adopt, for convenience, the standard *potential outcome* (or *counterfactual*) framework: for individual i, the observed outcome is related to the two possible potential outcomes $(\mathbf{Y}_{0i}, \mathbf{Y}_{1i})$ by $Y_i = (1 - Z_i)\mathbf{Y}_{0i} + Z_i\mathbf{Y}_{1i}$. We assume that X_i includes a sufficient set of covariates to control for confounding in the sense that $Z_i \perp\!\!\!\perp (\mathbf{Y}_{0i}, \mathbf{Y}_{1i}) \,|\, X_i$ (cf. ignorable treatment assignment, Rosenbaum and Rubin, 1983, p. 43). The *propensity score* $e(X_i) \equiv \mathrm{pr}(Z_i = 1 \,|\, X_i)$ has the balancing property $Z_i \perp\!\!\!\perp X_i \,|\, e(X_i)$, which also implies that $(\mathbf{Y}_{0i}, \mathbf{Y}_{1i}) \perp\!\!\!\perp Z_i \,|\, e(X_i)$ – this scalar score is therefore useful in controlling for confounding.

If the covariate space X_i is high-dimensional, in practice the task of controlling for confounding often involves some covariate selection; here one can either select for the features that predict the outcome, or the treatment assignment. To represent this, let S_i and B_i represent some a priori selected subsets of the all observed features X_i, so that the latter can be partitioned as $X_i = (S_i, R_i)$ or $X_i = (B_i, C_i)$, where possibly $R_i = \emptyset$ and $C_i = \emptyset$. If the selected set of features S_i captures all relevant prognostic information, then $\mathbf{Y}_{0i} \perp\!\!\!\perp X_i \,|\, S_i$ (Hansen, 2008). For the remainder of the paper, we consider the stronger condition (i) $(\mathbf{Y}_{0i}, \mathbf{Y}_{1i}) \perp\!\!\!\perp X_i \,|\, S_i$, which requires that S_i also captures all relevant information about possible effect modification.

In this case S_i is sufficient to control for confounding, since from the properties of conditional independence (Dawid, 1979) it follows that $(\mathbf{Y}_{0i}, \mathbf{Y}_{1i}) \perp\!\!\!\perp Z_i \,|\, S_i$. If, on the other hand, the selected set of features B_i has the balancing property (ii) $Z_i \perp\!\!\!\perp X_i \,|\, B_i$, it is sufficient to control for confounding. We are interested in estimation procedures which are valid when either $(\mathbf{Y}_{0i}, \mathbf{Y}_{1i}) \perp\!\!\!\perp X_i \,|\, S_i$ (but possibly $Z_i \not\perp\!\!\!\perp X_i \,|\, B_i$) or $Z_i \perp\!\!\!\perp X_i \,|\, B_i$ (but possibly $(\mathbf{Y}_{0i}, \mathbf{Y}_{1i}) \not\perp\!\!\!\perp X_i \,|\, S_i$).

Our parameter of interest is an average causal contrast such as

$$\mathrm{E}(\mathbf{Y}_{1i}) - \mathrm{E}(\mathbf{Y}_{0i}) = \mathrm{E}_{X_i}\{\mathrm{E}(\mathbf{Y}_{1i} \,|\, X_i)\} - \mathrm{E}_{X_i}\{\mathrm{E}(\mathbf{Y}_{0i} \,|\, X_i)\}.$$

Interest then lies in the identifiability of the average potential outcomes based on the observed data.

When the 'no unmeasured confounder' assumption holds, it follows that (e.g. Hernán and Robins, 2006, p. 43)

$$\mathrm{E}(\mathbf{Y}_{1i}) - \mathrm{E}(\mathbf{Y}_{0i}) = \int_{x_i} \{\mathrm{E}(Y_i \mid Z_i = 1, x_i) - \mathrm{E}(Y_i \mid Z_i = 0, x_i)\}\, p(x_i)\, \mathrm{d}x_i.$$

2. Bayesian model formulation for outcome regression

Any Bayesian model specification is constructed via de Finetti's representation for exchangeable random sequences $v_i \equiv (x_i, y_i, z_i)$ (see for example Bernardo and Smith, 1994, Chap. 4). The (subjective) joint distribution for observations $v \equiv (v_1, \ldots, v_n)$ may then be represented by

$$p(v) = \int_{\phi,\gamma,\psi} \left\{ \prod_{i=1}^{n} p(y_i \mid z_i, x_i; \phi) p(z_i \mid x_i; \gamma) p(x_i; \psi) \right\} \pi_0(\phi, \gamma, \psi)\, \mathrm{d}\phi\, \mathrm{d}\gamma\, \mathrm{d}\psi, \tag{1}$$

implying the existence of the parametric models and the prior density $\pi_0(\phi, \gamma, \psi)$. Since the representation theorem is not constructive, that is, does not specify the models implicit in (1), in practice inferences about a given finite-dimensional parametrization involves the often implicit assumption that $p(y_i \mid z_i, x_i; \phi_0) = f(y_i \mid z_i, x_i)$, $p(z_i \mid x_i; \gamma_0) = f(z_i \mid x_i)$ and $p(x_i; \psi_0) = f(x_i)$, where $(\phi_0, \gamma_0, \psi_0)$ is the limiting value of the posterior $\pi_n(\phi, \gamma, \psi) \equiv p(\phi, \gamma, \psi \mid v)$ in the sense of e.g. van der Vaart (1998, p. 139), and where the fs represent the 'true' limiting (sampling) distributions. We might further assume that the parameters are a priori independent, so that $\pi_0(\phi, \gamma, \psi) = \pi_0(\phi)\pi_0(\gamma)\pi_0(\psi)$, in which case it it also follows also that the posterior factorizes as $\pi_n(\phi, \gamma, \psi) = \pi_n(\phi)\pi_n(\gamma)\pi_n(\psi)$ (e.g. Gelman et al., 2004, p. 354–355).

The marginal distribution $p(x_i; \psi)$ can in practice be specified nonparametrically and estimated using the empirical covariate distribution, leading to a Bayesian estimator of the average causal contrast,

$$\int_{\phi} \frac{1}{n} \sum_{i=1}^{n} \{m(1, x_i; \phi) - m(0, x_i; \phi)\}\, \pi_n(\phi)\, \mathrm{d}\phi, \tag{2}$$

where $\pi_n(\phi) \propto \prod_{i=1}^{n} p(y_i \mid z_i, x_i; \phi)\pi_0(\phi)\, \mathrm{d}\phi$, and $m(z, x; \phi) \equiv \int y p(y \mid z, x; \phi)\, \mathrm{d}y$. In Supplementary Appendix 1, we show that (2) can be motivated without the use of potential outcomes notation through posterior predictive expectations for a new observation under a hypothetical completely randomized setting.

The estimator (2) is the Bayesian version of the well-known direct standardization or g-formula. We return to it in Section 6, but note here that it does not feature the treatment assignment model; rather, the posterior predictive approach for estimating the marginal causal contrast depends entirely on correct specification of the distribution $p(y_i \mid z_i, x_i; \phi)$, or in a moment-based representation, $m(z, x; \phi)$. Before discussing Bayesian alternatives to (2), we briefly review some commonly used frequentist approaches for combining outcome regression and propensity score adjustment.

3. Frequentist approaches for combining outcome regression and propensity score adjustment

3. Including propensity scores into outcome regression

Because of the balancing property of the propensity score, it is tempting to specify a propensity score $e(b_i; \gamma) \equiv \mathrm{pr}(Z_i = 1 \mid b_i; \gamma)$ and use a statistical model such as $p\{y_i \mid z_i, e(b_i), s_i; \phi\}$, in the hope that, if the prognostic model is is misspecified, adjusting for the propensity score would still sufficiently control for any residual confounding. For simplicity, we take the parameters ϕ to specify also the functional dependence between the propensity score and the outcome; to model this dependency, it is advisable to use flexible formulations such as splines (e.g. Zhang and Little, 2009). Using such an outcome model, the marginal causal contrast would then be estimated by

$$\frac{1}{n} \sum_{i=1}^{n} \left[m\left\{1, e(b_i; \widehat{\gamma}), s_i; \widehat{\phi}\right\} - m\left\{0, e(b_i; \widehat{\gamma}), s_i; \widehat{\phi}\right\} \right], \tag{3}$$

where $m\{z, e(b; \gamma), s; \phi\} \equiv \int y p\{y \mid z, e(b; \gamma), s; \phi\}\, \mathrm{d}y$, and where $\widehat{\phi}$ and $\widehat{\gamma}$ are maximum likelihood estimators for the parameters in the outcome regression and treatment assignment model, respectively. The motivation for such a double adjustment is that it is sufficient to control for confounding if either

condition (i) or (ii) of Section 2.1 applies. We summarize this property in the following theorem (a proof in Supplementary Appendix 2; in the notation all convergencies are in probability unless otherwise stated).

THEOREM 31. *Estimator* (3) *is consistent if the outcome model is correctly specified in the sense that* $m\{z, e(b;\gamma_0), s;\phi_0\} = \int y f(y \mid z, e(b), s)\, dy$*, the parameters in this can be consistently estimated so that* $\widehat{\phi} \to \phi_0$*, the treatment assignment model is correctly specified in the sense that* $p(z_i \mid b_i;\gamma_0) = f(z_i \mid b_i)$ *and* $\widehat{\gamma} \to \gamma_0$*, and either (i) holds true, or (ii) holds true.*

The estimator (3) may be considered 'doubly robust' in terms of the covariate selection in the sense that only one of the sets S_i and B_i needs to be correctly specified, although it still always relies on a correct parametric specification of the model for the expected outcome, conditional on $\{Z_i, e(B_i), S_i\}$.

3. The clever covariate and augmented outcome regression

The estimator discussed in the previous section did not specify which function of the propensity score $e(B_i)$ should be added to the regression model. Scharfstein et al. (1999, p. 1141–1142) and Bang and Robins (2005, p. 964–965) drew a connection between propensity score regression adjustment and doubly robust estimators of the form

$$\frac{1}{n}\sum_{i=1}^{n}\frac{y_i z_i - \{z_i - e(b_i;\widehat{\gamma})\}m(1, s_i;\widehat{\phi})}{e(b_i;\widehat{\gamma})} - \frac{1}{n}\sum_{i=1}^{n}\frac{y_i(1-z_i) - [(1-z_i) - \{1-e(b_i;\widehat{\gamma})\}]m(0, s_i;\widehat{\phi})}{1-e(b_i;\widehat{\gamma})},$$

which can be equivalently represented as

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{z_i}{e(b_i;\widehat{\gamma})} - \frac{1-z_i}{1-e(b_i;\widehat{\gamma})}\right\}\left\{y_i - m(z_i, s_i;\widehat{\phi})\right\} + \frac{1}{n}\sum_{i=1}^{n}\left\{m(1, s_i;\widehat{\phi}) - m(0, s_i;\widehat{\phi})\right\}. \quad (4)$$

On considering the score equation derived from a regression of Y_i on Z_i and S_i with mean function $m(z, s;\phi)$, this form suggests incorporating the derived covariate $c(z_i, b_i) = z_i/e(b_i) - (1-z_i)/\{1 - e(b_i)\}$ (termed the 'clever covariate' by Rose and van der Laan, 2008, p. 8) additively into the outcome regression, that is, for example

$$m(z, s;\phi) = \phi_0 + \phi_1 z + \phi_2^{\top} s + \phi_3 c(z, b). \quad (5)$$

The first term in (4) is then zero through the maximum likelihood score equation, leaving only the last term which is the model based estimator of the marginal treatment effect. Thus with the clever covariate in the outcome model, the doubly robust estimator is equivalent to the model-based estimator. In the special case of model (5), this becomes

$$\frac{1}{n}\sum_{i=1}^{n}\left\{m(1, x_i;\widehat{\phi}) - m(0, x_i;\widehat{\phi})\right\} = \widehat{\phi}_1 + \widehat{\phi}_3\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{e(b_i;\widehat{\gamma})} - \frac{1}{1-e(b_i;\widehat{\gamma})}\right\}.$$

A potential drawback of using this covariate is that it may lead to extreme variability for the resulting mean difference estimator, even compared to inverse probability of treatment weighted estimators of the form

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i z_i}{e(b_i;\widehat{\gamma})} - \frac{y_i(1-z_i)}{1-e(b_i;\widehat{\gamma})}\right\}. \quad (6)$$

To see why this is the case, the distribution of $c(z_i, b_i)$ in itself can be very skewed to the right, but this becomes even more pronounced in the model-based estimator where the clever covariate has to evaluated at both $c(1, b_i)$ and $c(0, b_i)$ for each $i = 1, \ldots, n$. In contrast, the inverse probability of treatment weighted estimator (6) involves only the probabilities of treatments that were actually assigned.

Eq. (4) is doubly robust in a stronger, semi-parametric, sense than (3); it does not require correct specification of the outcome model, if the treatment assignment model is correctly specified. The approximate Bayesian double robust approach proposed by Graham et al. (2015) involved replacing $m(z, x;\phi)$

in (2) with a linear predictor augmented with the clever covariate. We take this to be a special case of the two-step Bayesian methods to be discussed in Section 4, and thus do not separately consider it in the present paper. However, in Section 6.2 we will show how the form (4) may be derived through posterior predictive expectations and importance sampling.

3. Inverse probability of treatment weighted outcome regression

Yet another estimator for the marginal causal contrast is

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\mathrm{E}(\mathbf{Y}_{1i}\,|\,s_i;\widehat{\phi}) - \mathrm{E}(\mathbf{Y}_{0i}\,|\,s_i;\widehat{\phi})\right\}, \tag{7}$$

where the parameters ϕ in the model for the potential outcomes $(\mathbf{Y}_{1i}, \mathbf{Y}_{0i})$ are estimated using an inverse probability of treatment weighted estimating function $l(\phi) = \sum_{i=1}^{n} l_i(\phi)$, where

$$l_i(\phi) = \sum_{a=0}^{1} \mathbf{1}_{\{z_i=a\}} \frac{\log p(\mathbf{y}_{ai}\,|\,s_i;\phi)}{\mathrm{pr}(Z_i = a\,|\,b_i)}.$$

The corresponding estimating equation is $u(\phi) = \sum_{i=1}^{n} u_i(\phi) = 0$, where the pseudo-score function is

$$u_i(\phi) = \sum_{a=0}^{1} \mathbf{1}_{\{z_i=a\}} \frac{\frac{\partial}{\partial\phi}\log p(\mathbf{y}_{ai}\,|\,s_i;\phi)}{\mathrm{pr}(Z_i = a\,|\,b_i)}. \tag{8}$$

Here the treatment assignment probabilities $\mathrm{pr}(Z_i = a\,|\,b_i)$ would in practice be replaced with estimates $\mathrm{pr}(Z_i = a\,|\,b_i;\widehat{\gamma})$. To motivate the use of (8) for the parameter estimation, we present the following theorem (a proof given in Supplementary Appendix 2).

THEOREM 32. *The estimating equation $u(\phi) = 0$ is unbiased if the model for the potential outcomes is correctly specified in the sense that $p(\mathbf{y}_{ai}\,|\,s_i;\phi_0) = f(\mathbf{y}_{ai}\,|\,s_i)$, and either (i) or (ii) holds true.*

If we further assume the consistency of the estimator for ϕ, as well as consistency of $\widehat{\gamma}$ when the weights are correctly specified, it also follows that (7) consistently estimates the marginal causal contrast. In Section 6.1 we demonstrate that an estimator of the form (7) can also be motivated from Bayesian arguments.

4. Two-step estimation when the propensity score is unknown

In observational settings the function $e(B_i)$ is unknown and has to be estimated. When using an estimator of the form (3), a central question from a Bayesian perspective then is how the uncertainty in the estimation of the parameters γ is incorporated in the inference of the marginal causal contrast. A 'Bayesian' approach could be motivated by writing the posterior predictive expectation as

$$\int_{\psi,\gamma,\phi}\int_{x_i} m(a, e(b_i;\gamma), s_i;\phi)p(x_i;\psi)\pi_n(\phi\,|\,\gamma)\pi_n(\gamma)\pi_n(\psi)\,\mathrm{d}x_i\,\mathrm{d}\phi\,\mathrm{d}\gamma\,\mathrm{d}\psi, \tag{9}$$

where

$$\pi_n(\phi\,|\,\gamma) \propto \prod_{i=1}^{n} p\{y_i\,|\,z_i, e(b_i;\gamma), s_i;\phi\}\pi_0(\phi) \quad \text{and} \quad \pi_n(\gamma) \propto \prod_{i=1}^{n} p(z_i\,|\,b_i;\gamma)\pi_0(\gamma).$$

The integrals of the form (9) could be evaluated by Monte Carlo integration, by forward sampling first from $\pi_n(\gamma)$ and given the current value γ, from the conditional posterior $\pi_n(\phi\,|\,\gamma)$. However, a concern related to such an approach is that the product of the posterior distributions $\pi_n(\phi\,|\,\gamma)$ and $\pi_n(\gamma)$ in (9) does not necessarily correspond to any well defined joint posterior, except in the case when the outcome is correctly specified in the sense (i), in which case it does not depend on the propensity score. In contrast, for the correctly specified models in (1) we indeed have that $p(\phi\,|\,v)p(\phi\,|\,x,z) = p(\phi,\gamma\,|\,v)$. As a result, the above outlined two-step approach is not proper Bayesian, and would have to be evaluated on its frequency-based properties.

To give an example of a situation where the two-step Bayesian approach does not result in correct frequency-based inferences, we can consider the special case of the outcome model $m\{z, e(b;\gamma), s;\phi\} =$

$\phi_0 + \phi_1 z + \phi_2^\top s + \phi_3^\top g\{e(b;\gamma)\}$, where g is, for example, some appropriate spline basis transformation of the propensity score. With this model the estimator based on (9) for the average causal contrast $\mathrm{E}(\mathbf{Y}_{1i}) - \mathrm{E}(\mathbf{Y}_{0i})$ reduces to $\int_{\phi,\gamma} \phi_1 \pi_n(\phi \mid \gamma)\pi_n(\gamma)\,\mathrm{d}\phi\,\mathrm{d}\gamma$, that is, to an estimator of the posterior mean $\mathrm{E}_{\Gamma \mid X,Z}\{\mathrm{E}(\Phi_1 \mid v;\gamma)\}$. This estimator in turn can be approximated with $\sum_{j=1}^m \widehat{\phi}_1(\gamma^{(j)})/m$, where $(\gamma^{(1)},\ldots,\gamma^{(m)})$ is a Monte Carlo sample from $\pi_n(\gamma)$ (cf. Kaplan and Chen, 2012, p. 589). We note first that this estimator has the same asymptotic distribution as $\widehat{\phi}_1(\widehat{\gamma})$, where the treatment assignment model parameters have been fixed to their maximum likelihood estimates (see Supplementary Appendix 3). In Supplementary Appendix 3 we further show that $\mathrm{avar}\{\widehat{\phi}_1(\widehat{\gamma})\} \leq \mathrm{avar}\{\widehat{\phi}_1(\gamma_0)\}$, where $\widehat{\phi}_1(\gamma_0)$ is the estimator given the 'true' propensity scores. Thus, with the propensity score adjusted outcome model specification, a variance adjustment due to estimating the propensity scores should reduce the asymptotic variance of the resulting treatment effect estimator compared to a hypothetical situation where the true propensity scores are known (cf. Henmi and Eguchi, 2004). In contrast, Kaplan and Chen (2012, p. 592) and Graham et al. (2015, p. 11) propose variance estimators based on the variance decomposition formula

$$\mathrm{var}(\Phi_1 \mid v) = \mathrm{E}_{\Gamma \mid X,Z}\left\{\mathrm{var}(\Phi_1 \mid v;\gamma)\right\} + \mathrm{var}_{\Gamma \mid X,Z}\left\{\mathrm{E}(\Phi_1 \mid v;\gamma)\right\}, \tag{10}$$

which appears to add a further variance component. An explanation for the discrepancy is that with the correctly specified models in the representation (1), we have that $p(\phi_1 \mid v;\gamma) = p(\phi_1 \mid v)$, and the the second variance component becomes zero. On the other hand, if the models are misspecified, the product form likelihood in the representation (1) does not apply in the first place. This illustrates the difficulty in applying Bayesian procedures in the context of incompatible models. Even though this is routinely done in the context of multiple imputation (e.g. Rubin, 1996), and often produces reasonable results, in the present context there is little motivation to use an approach which introduces an additional variance component to the posterior variance, given that estimation of the propensity scores should reduce the variance of the treatment effect estimator. We further discuss this discrepancy in the following section.

5. JOINT ESTIMATION OF OUTCOME AND TREATMENT ASSIGNMENT MODELS

Even though the outcome Y_i can obviously be predictive of the individual treatment assignment Z_i, the outcomes are not informative of the treatment assignment mechanism (a proof in Supplementary Appendix 2):

THEOREM 53. *If the outcome model is correctly specified, the outcomes are non-informative of the parameters characterizing the treatment assignment process.*

In such a case the treatment assignment model plays no part in the inferences, since the corresponding posterior predictive estimator is (2). However, the Bayesian propensity score approach proposed by McCandless et al. (2009a) specifies a parametrization making the outcome and treatment assignment models dependent and estimates the parameters jointly. More recently, Zigler et al. (2013) suggested that a similar approach could be used to obtain a Bayesian analogue to doubly robust inferences. Such an approach can be understood by assuming that there exists a de Finetti parametrization (ϕ^*,γ^*) for which

$$p(v) = \int_{\phi^*,\gamma^*,\psi} \left\{\prod_{i=1}^n p(y_i, z_i \mid x_i;\phi^*,\gamma^*)p(x_i;\psi)\right\} \pi_0(\phi^*)\pi_0(\gamma^*)\pi_0(\psi)\,\mathrm{d}\phi^*\,\mathrm{d}\gamma^*\,\mathrm{d}\psi,$$

where $p(y_i, z_i \mid x_i;\phi^*,\gamma^*) = p\{y_i \mid z_i, e(b_i;\gamma^*), s_i;\phi^*\}p(z_i \mid b_i;\gamma^*)$. Compared to (ϕ,γ) in (1), neither ϕ^* or γ^* retains the original interpretation, but now there is a well defined joint posterior distribution

$$\pi_n(\phi^*,\gamma^*) \propto \prod_{i=1}^n \left[p\{y_i \mid z_i, e(b_i;\gamma^*), s_i;\phi^*\}p(z_i \mid b_i;\gamma^*)\right] \pi_0(\phi^*)\pi_0(\gamma^*).$$

Inferences could now be based on the posterior predictive expectations

$$\int_{\phi^*,\gamma^*,\psi} \int_{x_i} m\{a, e(b_i;\gamma^*), s_i;\phi^*\}p(x_i;\psi)\pi_n(\phi^*,\gamma^*)\pi_n(\psi)\,\mathrm{d}x_i\,\mathrm{d}\phi^*\,\mathrm{d}\gamma^*\,\mathrm{d}\psi. \tag{11}$$

At first sight, (11) would seem more natural than (9), since the specification (11) does not make use of incompatible models. However, now the quantities $e(b_i;\gamma^*)$ do not possess the balancing properties of

propensity scores, and thus it would be difficult to show whether (11) would have the 'double robustness' property of the estimator (3).

To address the lack of balance, McCandless et al. (2010) suggested a Gibbs sampler type approach similar to that of Lunn et al. (2009) to cut the feedback from the outcome model, by successively drawing from the conditional posteriors $\pi_n(\gamma)$ and $\pi_n(\phi \mid \gamma)$ to approximate the joint posterior of (ϕ^*, γ^*). However, as discussed in the previous Section, these posteriors are incompatible and generally such a sampling procedure is not guaranteed to converge to any well defined joint distribution. In fact, if the conditional posteriors can be sampled directly, or if the second sampling step is allowed to converge to the corresponding conditional distribution, the inferences based on the formulations (9) and (11) will be equivalent.

To sum up, trying to recover fully probabilistic inferences through sampling from a joint posterior of the outcome and treatment assignment model parameters loses the balancing property of the propensity scores, and consequently, the properties of the resulting estimator would be difficult to establish. On the other hand, cutting the feedback in an attempt to recover the balancing property would mean that the inferences are no longer based on well-defined posterior distributions. Thus, in the following section, following the approach outlined in Saarela et al. (2015b), we formulate alternative Bayesian estimators that are not based on Bayesian propensity score adjustment.

6. Posterior predictive inferences with importance sampling

6. *Inverse probability of treatment weighted outcome regression*

It has been recognized by various authors (e.g. Røysland, 2011; Chakraborty and Moodie, 2013) that inverse probability of treatment weighting can be motivated through a change of probability measures, or equivalently, importance sampling. However, as far as we know, before Saarela et al. (2015b) this approach has not been used to formulate Bayesian causal inferences. Here we argue that this approach can be used to resolve the paradoxes discussed in Sections (4) and (5). We follow the posterior predictive reasoning of Supplementary Appendix 1, but rather than trying to directly predict a new observation under the experimental setting $\mathcal{E}$, we consider first the task of parameter estimation under this regime. For this purpose, a Bayes estimator for the outcome model parameters can be constructed by maximizing the expected utility $\mathrm{E}_{\mathcal{E}}\{l(\phi; V_i) \mid v\}$ with respect to ϕ, where the log-likelihood $l(\phi; v_i) \equiv \log p(y_i \mid z_i, s_i; \phi)$ takes the role of a parametric utility function, and the expectation is over a predicted new observation $v_i = (y_i, z_i, x_i)$, $i \notin \{1, \ldots, n\}$.

Let further ξ be a set of parameters characterizing the entire data-generating mechanism under the observational regime $\mathcal{O}$. We can further write the expectation as

$$\mathrm{E}_{\mathcal{E}}\{l(\phi; V_i) \mid v\} = \mathrm{E}_{\Xi \mid V}\left[\mathrm{E}_{\mathcal{E}}\{l(\phi; V_i) \mid v; \xi\}\right],$$

where, following Walker (2010, p. 26–27), we can consider the lower dimensional decision of maximizing the expected utility $\mathrm{E}_{\mathcal{E}}\{l(\phi; V_i) \mid v; \xi\}$ with respect to ϕ conditional on ξ. With a known regime $\mathcal{E}$ and the stability assumption discussed in Supplementary Appendix 1, $\arg\max_\phi \mathrm{E}_{\mathcal{E}}\{l(\phi; V_i) \mid v; \xi\}$ is a deterministic function of ξ. Thus, the uncertainty represented by the posterior distribution $p_{\mathcal{O}}(\xi \mid v)$ then also reflects the uncertainty on ϕ, providing means to construct a posterior distribution for ϕ. This proceeds as follows; the inner expectation can be written as

$$\begin{aligned}
\mathrm{E}_{\mathcal{E}}[l(\phi; V_i) \mid v; \xi] &= \int_{v_i} l(\phi; v_i) p_{\mathcal{E}}(v_i \mid v; \xi)\, \mathrm{d}v_i \\
&= \int_{v_i} l(\phi; v_i) \frac{p_{\mathcal{E}}(v_i \mid v; \xi)}{p_{\mathcal{O}}(v_i \mid v; \xi)} p_{\mathcal{O}}(v_i \mid v; \xi)\, \mathrm{d}v_i \\
&= \int_{v_i} l(\phi; v_i) \frac{p_{\mathcal{E}}(y_i \mid z_i, x_i, v; \xi) p_{\mathcal{E}}(z_i) p_{\mathcal{E}}(x_i \mid v; \xi)}{p_{\mathcal{O}}(y_i \mid z_i, x_i, v; \xi) p_{\mathcal{O}}(z_i \mid x_i, v; \xi) p_{\mathcal{O}}(x_i \mid v; \xi)} p_{\mathcal{O}}(v_i \mid v; \xi)\, \mathrm{d}v_i \\
&= \int_{v_i} l(\phi; v_i) \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}(z_i \mid x_i, v; \xi)} p_{\mathcal{O}}(v_i \mid v; \xi)\, \mathrm{d}v_i,
\end{aligned}$$

where in the last equality we made use of the stability assumption of Supplementary Appendix 1. In the last form we can replace the predictive distribution under $\mathcal{O}$ with the Bayesian bootstrap specification $p_{\mathcal{O}}(v_i \mid v; \xi) = \sum_{k=1}^{n} \xi_k \delta_{v_k}(v_i)$, where $\xi \equiv (\xi_1, \ldots, \xi_n)$ and $\Xi \mid v \sim \mathrm{Dirichlet}(1, \ldots, 1)$, as in the weighted likelihood bootstrap of Newton and Raftery, 1994. Denoting $w_i(\xi) \equiv p_{\mathcal{E}}(z_i)/p_{\mathcal{O}}(z_i \mid x_i, v; \xi)$,

the expected utility now becomes

$$\mathrm{E}_{\mathcal{E}}[l(\phi; V_i) \mid v; \xi] = \int_{v_i} l(\phi; v_i) w_i(\xi) \sum_{k=1}^{n} \xi_k \delta_{v_k}(v_i) \, \mathrm{d}v_i = \sum_{k=1}^{n} w_k(\xi) \xi_k l(\phi; v_k), \tag{12}$$

that is, a weighted log-likelihood, motivating the estimator

$$\widehat{\phi}(\xi) \equiv \arg\max_{\phi} \sum_{k=1}^{n} w_k(\xi) \xi_k l(\phi; v_k).$$

An approximate posterior distribution for ϕ under $\mathcal{E}$ can now be constructed by repeatedly sampling the weight vectors from $\Xi \mid v \sim \text{Dirichlet}(1, \ldots, 1)$, and recalculating $\widehat{\phi}(\xi)$ at each realization. This approach of creating a mapping between the non-parametric specification and a parametrization relevant to inferences is analogous to Chamberlain and Imbens (2003), but adds the importance sampling weights to the Dirichlet weights in order to make inferences across the observational and experimental regimes.

The weights $w_i(\xi)$ function as importance sampling weights in Monte Carlo integration; they add variability to the estimation, but in the present context provide some protection towards misspecification of the outcome model, in the sense of Section 3.3. The above did not yet address how to calculate these weights; in principle these are fully determined by the current realization of ξ under the non-parametric specification, but in practice parametric model specifications are needed for smoothing purposes, and we need a way to link the ξ and the treatment assignment model parameters γ. For this purpose γ itself can be estimated through the weighted likelihood bootstrap since this readily gives the deterministic function linking the two parametrizations; thus in (12) we choose $w_i(\xi) = p_{\mathcal{E}}(z_i)/p_{\mathcal{O}}\{z_i \mid b_i; \widehat{\gamma}(\xi)\}$, where $\widehat{\gamma}(\xi) \equiv \arg\max_{\gamma} \sum_{k=1}^{n} \xi_k \log p(z_k \mid b_k; \gamma)$. The probabilities $p_{\mathcal{E}}(z_i)$ are given by the chosen regime $\mathcal{E}$ that is the object of inference; in practice the estimation is most efficient when we choose the target regime to be as close as possible to the observed regime $\mathcal{O}$; this can be achieved by fixing $p_{\mathcal{E}}(z_i)$ to the marginal treatment assignment probabilities under $\mathcal{O}$, which would result in the usual kind of stabilized inverse probability of treatment weights used in marginal structural modelling (Robins et al., 2000; Hernán et al., 2001; Cole and Hernán, 2008).

The marginal causal contrast may now be estimated through the expectations

$$\begin{aligned} \mathrm{E}_{\Xi \mid V}\{\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = a, v; \xi)\} &= \int_{\xi} \int_{s_i} m\{a, s_i; \widehat{\phi}(\xi)\} \sum_{k=1}^{n} \xi_k \delta_{s_k}(s_i) p_{\mathcal{O}}(\xi \mid v) \, \mathrm{d}s_i \, \mathrm{d}\xi \\ &= \int_{\xi} \sum_{k=1}^{n} \xi_k m\{a, s_k; \widehat{\phi}(\xi)\} p_{\mathcal{O}}(\xi \mid v) \, \mathrm{d}\xi, \end{aligned} \tag{13}$$

where we used the non-parametric specification $p(s_i \mid v; \xi) = \sum_{k=1}^{n} \xi_k \delta_{s_k}(s_i)$, and where again $p_{\mathcal{O}}(\xi \mid v)$ is replaced with the uniform Dirichlet distribution. Since all uncertainty is contained in the parameter vector ξ, a posterior distribution for the predictive mean or mean difference can be constructed as before through resampling. The point estimator given by (13) is the direct Bayesian analogue of (7), where the outcome model was estimated using inverse probability of treatment weighted regression. In fact, if we fix $\xi_k = 1/n$, $k = 1, \ldots, n$, instead of considering these as unknown parameters, the two estimators are equivalent. Thus, we conjecture that the estimator given by (13) has a similar 'double robustness' property as (7). We demonstrate this through simulations in Section 7, but before that, we show how the semi-parametric doubly robust estimator (4) can be motivated as a posterior predictive expectation.

6. *Doubly robust estimation*

In Supplementary Appendix 4 we show that under the non-parametric specification in terms of ξ, the posterior predictive causal contrast may be written as

$$\begin{aligned} &\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 1, v; \xi) - \mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 0, v; \xi) \\ &\quad = \sum_{k=1}^{n} \xi_k \{y_i - m(z_k, x_k; \xi)\} \left\{ \frac{z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \mid x_k, v; \xi)} - \frac{1 - z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 0 \mid x_k, v; \xi)} \right\} \\ &\quad\quad + \sum_{k=1}^{n} \xi_k \left\{ m(1, x_k; \xi) - m(0, x_k; \xi) \right\}, \end{aligned} \tag{14}$$

which corresponds to formulation (4). Since the non-parametric specification places no restrictions on the conditional distributions, in practice, to obtain an estimator, the non-parametrically specified quantities $m(z_k, x_k; \xi)$ and $\mathrm{pr}_{\mathcal{O}}(Z_k = a \,|\, x_k, v; \xi)$ would have to be replaced with the parametric versions $m\{z_k, s_k; \widehat{\phi}(\xi)\}$ and $\mathrm{pr}_{\mathcal{O}}\{Z_k = a \,|\, b_k; \widehat{\gamma}(\xi)\}$, estimated using the weighted likelihood bootstrap, as in the previous section. It is straightforward to see that if the outcome model is correctly specified, the expression (14) reduces to the model-based estimator (the second additive term). This again reflects the fact that if we believe in our outcome model, the treatment assignment model does not play a part in the inferences. However, a Bayesian might want to use an estimator of the form (14) if being restricted by two parametric constraints, in terms of ϕ and γ, but not knowing which one of these is correct. If either one of the parametric specifications is correct, (14) will still reduce to the posterior predictive mean difference, the natural Bayesian estimator. We summarize this property in the following theorem (a proof in Supplementary Appendix 2).

THEOREM 64. *The estimator obtained by substituting the parametric specifications* $m\{z_k, s_k; \widehat{\phi}(\xi)\}$ *and* $\mathrm{pr}_{\mathcal{O}}\{Z_k = a \,|\, b_k; \widehat{\gamma}(\xi)\}$ *into expression* (14) *is equivalent to the posterior predictive mean difference if either one of these models is correctly specified.*

A posterior distribution for the mean difference can be generated as in the previous section through resampling of the parameter vectors ξ and recalculating (14) at each realization; we will illustrate this in the following section.

7. SIMULATION STUDY

Above we have made a distinction between model misspecification due to omission of relevant covariates, and misspecification of the parametric functional relationship between the outcome and the covariates, and noted that all the estimators discussed in Section 3 should be 'doubly robust' against the former type of misspecification. However, in practice the consequences of these two types of misspecification will often be similar; they result in residual confounding. Therefore, herein we investigate how the different estimators perform when the covariate sets S_i and B_i are not only created by a partitioning, but also a transformation of the x_i's. For this purpose, we simulated $X_{ij} \sim \mathrm{N}(0, 1)$, $j = 1, \ldots, 4$, and transformed these as $c_{ij} = |x_{ij}|/(1 - 2/\pi)^{1/2}$. The true treatment assignment and outcome mechanisms were specified as $Z_i \,|\, x_i \sim \mathrm{Bernoulli}(\mathrm{expit}\{0.4c_{i1} + 0.4x_{i2} + 0.8x_{i3}\})$ and $Y_i \,|\, z_i, x_i \sim \mathrm{N}(z_i - c_{i1} - x_{i2} - x_{i4}, 1)$, respectively. For estimation, we considered two scenarios: (I) $s_i \equiv (x_{i1}, x_{i2}, x_{i3})$ and $b_i \equiv (c_{i1}, x_{i2}, x_{i4})$ (misspecified outcome model and correctly specified treatment assignment model), and (II) $s_i \equiv (c_{i1}, x_{i2}, x_{i3})$ and $b_i \equiv (x_{i1}, x_{i2}, x_{i4})$ (correctly specified outcome model and misspecified treatment assignment model).

We are interested in the marginal causal contrast $\mathrm{E}(\mathbf{Y}_{i1}) - \mathrm{E}(\mathbf{Y}_{i0}) = 1$. To estimate this, we applied the Bayesian estimators discussed in Sections 4, 5, and 6. In the two-step estimation we attempted both forward sampling from the posterior distributions, and the variance decomposition formula (10). In the former, instead of Markov chain Monte Carlo, we applied normal approximations for the posterior distributions, of the form $\Phi \,|\, v; \gamma \sim \mathrm{N}\{\widehat{\phi}(\gamma), S\}$, where $\widehat{\phi}(\gamma)$ is the maximum likelihood estimate and S its estimated variance-covariance matrix. The posterior distribution $\Gamma \,|\, x, z$ was approximated using the weighted likelihood bootstrap. In the joint estimation, we again used a normal approximation, centered at the joint maximum likelihood estimates $(\widehat{\phi}, \widehat{\gamma})$, and the variance-covariance matrix given by the inverse of the observed information at the maximum likelihood point. In both two-step and joint estimation, the fitted models were specified as $m\{z_i, e(b_i; \gamma), s_i; \phi\} = \phi_0 + \phi_1 z_i + \phi_2^\top s_i + \phi_3^\top g\{e(b_i; \gamma)\}$, where g is a cubic polynomial basis, and $e(b_i; \gamma) = \mathrm{expit}(\gamma_0 + \gamma_1^\top b_i)$. In the importance sampling (IS) estimator proposed in Section 6.1, and in the importance sampling/doubly robust estimator (IS/DR) of Section 6.2, the fitted treatment assignment model was the same, with the outcome model specified through $m(z_i, s_i; \phi) = \phi_0 + \phi_1 z_i + \phi_2^\top s_i$.

For comparison to the Bayesian estimators, we also calculate naive unadjusted comparison (naive), outcome regression adjusted for covariates s_i (adjusted), the standard inverse probability of treatment weighted estimator (6) (IPTW), the semi-parametric doubly robust estimator (4) (DR), the clever covariate version of this (CC), the inverse probability of treatment weighted outcome regression based estimator (7) (OR/IPTW), as well as propensity score adjusted outcome regression based estimator (3) (OR/PS). For IPTW, DR, CC, and OR/IPTW, the standard errors were estimated through the standard frequentist nonparametric bootstrap (Efron, 1979). For OR/PS, to demonstrate the variance estimation

issues discussed in Section 4, we calculated both observed information based standard errors, and the adjusted sandwich type standard errors discussed in Supplementary Appendix 3.

The results over 1000 replications are shown in Table 1. Under scenario (I), all estimators except naive and adjusted can correct for confounding, although the joint estimation approach produces a slight bias. In terms of efficiency, the estimators based on propensity score adjusted outcome regression are the best, with the inverse probability of treatment weighting based estimators losing slightly. As discussed in Section 3.2, the clever covariate approach results in higher variability compared to the other doubly adjusted estimators. In terms of variance estimation, the comparison between the unadjusted and adjusted standard errors for OR/PS suggests that under this simulation setting estimation of the propensity scores substantially reduces the variance, and not adjusting for this results in overcoverage. The resampling based variance estimators adjust for this automatically. However, the two-step approach to variance estimation performs poorly; as demonstrated in Supplementary Appendix 3, the two-step point estimator has the same asymptotic variance as the other OR/PS estimators, but the two-step variance estimators unnecessarily add a further variance component. Thus, the simulation results support the discussion in Sections 4 and 5; the two-step and joint estimation approaches are difficult to justify from Bayesian arguments, and do not seem to provide practical advantages in terms of their frequency-based properties. On the other hand, the importance sampling based Bayesian estimators produce very similar results to the OR/IPTW and DR estimators, respectively.

Under scenario (II), all the estimators except IPTW are unbiased, which is expected based on their previously discussed theoretical properties. When the outcome model is correctly specified, there is also very little difference in the efficiency of the various estimators.

Table 1: Estimates for the marginal causal contrast (true value $= 1$) over 1000 simulation rounds

Scenario	Estimator	Point estimate	Relative bias (%)	SD	SE	MC error	Coverage (%)
(I)	Naive	0.347	−65.3	0.128	0.128	0.004	0.3
	Adjusted	0.667	−33.3	0.091	0.092	0.003	3.6
	IPTW	1.001	0.1	0.135	0.138	0.005	94.4
	OR/PS (obs. information)	0.997	−0.3	0.071	0.095	0.002	99.3
	OR/PS (adj. sandwich)	0.997	−0.3	0.071	0.073	0.002	95.3
	DR	0.998	−0.2	0.087	0.088	0.003	94.2
	Clever covariate	1.026	2.6	0.110	0.110	0.003	93.0
	OR/IPTW	0.990	−1.0	0.082	0.083	0.003	93.9
	Two-step (forward sampling)	0.999	−0.1	0.071	0.113	0.002	99.8
	Two-step (variance decomposition)	0.995	−0.5	0.071	0.112	0.002	99.8
	Joint estimation	1.046	4.6	0.071	0.071	0.002	89.5
	Importance sampling	0.991	−0.9	0.083	0.080	0.003	93.6
	Importance sampling/DR	0.997	−0.3	0.087	0.086	0.003	93.9
(II)	Naive	0.347	−65.3	0.128	0.128	0.004	0.3
	Adjusted	0.997	−0.3	0.065	0.067	0.002	95.3
	IPTW	0.629	−37.1	0.133	0.131	0.005	20.3
	OR/PS (obs. information)	0.997	−0.3	0.071	0.071	0.002	95.4
	OR/PS (adj. sandwich)	0.997	−0.3	0.071	0.071	0.002	95.4
	DR	0.999	−0.1	0.074	0.074	0.003	95.9
	Clever covariate	0.999	−0.1	0.075	0.075	0.003	95.9
	OR/IPTW	0.999	−0.1	0.074	0.073	0.003	95.7
	Two-step (forward sampling)	1.000	0.0	0.071	0.072	0.002	96.5
	Two-step (variance decomposition)	0.997	−0.3	0.071	0.071	0.002	95.5
	Joint estimation	0.997	−0.3	0.071	0.071	0.002	95.4
	Importance sampling	0.999	−0.1	0.074	0.072	0.003	95.2
	Importance sampling/DR	0.999	−0.1	0.074	0.073	0.003	95.5

The columns correspond to estimator, mean point estimate, relative bias, Monte Carlo standard deviation (SD), mean standard error estimate (SE), Monte Carlo (MC) error (batch means) of the mean point estimate, and 95% confidence interval coverage probability. The two scenarios correspond to (I) misspecified outcome model and correctly specified treatment assignment model, and (II) correctly specified outcome model and misspecified treatment assignment model.

8. DISCUSSION

In this paper we reviewed previously proposed Bayesian approaches for propensity score adjusted inferences, focusing on the assumptions concerning correct model specifications. Here it is important to make a distinction between misspecification due to omission of relevant covariates from the outcome model, and misspecification of the functional form of the dependency between the covariates and the outcome. The frequentist propensity score adjusted outcome regression is robust against the former type of model misspecification, but this property is lost in Bayesian estimation, if the misspecified outcome model is allowed to feed back to the estimation of the propensity scores. While feedback issue has been well documented in the literature (e.g. McCandless et al., 2009b; Zigler et al., 2013), and the reasons behind this were already stated by Robins and Ritov (1997), here we attempted to make the assumptions underlying the Bayesian propensity score approach more explicit. On the other hand, we point out that cutting this feedback in a two-step Bayesian estimation procedure unnecessarily inflates the posterior variance estimates.

As reaching double robustness through Bayesian propensity score adjustment looks difficult, herein we attempted a completely different approach through posterior predictive inferences. Our proposed approach decouples the outcome regression and treatment assignment model through introducing the inverse probability of treatment weights as importance sampling weights in Monte Carlo integration in evaluating posterior predictive expectations. A similar procedure was used in a marginal structural modelling context by Saarela et al. (2015b), improved to its present form in Saarela et al. (2015c). While they used the importance sampling approach for estimating marginal outcome models in a longitudinal setting, herein we showed that in a point treatment setting the combination of importance sampling and posterior predictive inferences can be used to motivate weighted outcome regression or semi-parametric doubly robust inferences. Such a possibility was mentioned, but not formally justified, by Saarela et al. (2015a) who applied the importance sampling procedure in the context of estimating optimal treatment regimes. The disadvantage of the importance sampling approach is the same as in the corresponding frequentist inverse probability of treatment weighted inference procedures: the importance sampling weights add variability to the point estimator. In order to control this, a standard approach would be to truncate the weights (e.g. Xiao et al., 2013), which would also be possible in the importance sampling context (Ionides, 2008). Recently, Vehtari & Gelman (2015, `arXiv:1507.02646v2`) suggested probabilistic truncation of importance sampling weights; studying this possibility in the present context is a topic for further research.

ACKNOWLEDGEMENT

The authors acknowledge support of the Natural Sciences and Engineering Research Council (NSERC) of Canada.

SUPPLEMENTARY MATERIAL

Supplementary material available includes Supplementary Appendices 1-4 referred to herein, containing proofs to theorems and other technical material.

APPENDIX 1

Causal inference as a prediction problem

The estimator (2) can be motivated without the use of potential outcomes notation as a posterior predictive expectation for a new observation under a hypothetical completely randomized setting $\mathcal{E}$ where $Z_i \perp\!\!\!\perp_{\mathcal{E}} X_i$ and the probabilities $\mathrm{pr}_{\mathcal{E}}(Z_i = a)$ are known constants (cf. the randomized trial measure discussed by Røysland, 2011). The data are observed under a setting $\mathcal{O}$, where $Z_i \not\perp\!\!\!\perp_{\mathcal{O}} X_i$, and causal

inference then corresponds to inference across these regimes. We can now write for $i \notin \{1, \ldots, n\}$

$$\begin{aligned}
&\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = a, v) \\
&\quad = \frac{\int_{\phi,\psi}\int_{x_i}\int_{y_i} y_i p(y_i \mid Z_i = a, x_i; \phi)\mathrm{pr}_{\mathcal{E}}(Z_i = a)p(x_i;\psi)\pi_n(\phi)\pi_n(\psi)\,\mathrm{d}y_i\,\mathrm{d}x_i\,\mathrm{d}\phi\,\mathrm{d}\psi}{\int_{\phi,\psi}\int_{x_i}\int_{y_i} p(y_i \mid Z_i = a, x_i; \phi)\mathrm{pr}_{\mathcal{E}}(Z_i = a)p(x_i;\psi)\pi_n(\phi)\pi_n(\psi)\,\mathrm{d}y_i\,\mathrm{d}x_i\,\mathrm{d}\phi\,\mathrm{d}\psi} \\
&\quad = \int_{\phi,\psi}\int_{x_i} m(a, x_k; \phi)p(x_i;\psi)\pi_n(\phi)\pi_n(\psi)\,\mathrm{d}x_i\,\mathrm{d}\phi\,\mathrm{d}\psi \qquad (15) \\
&\quad = \int_{\phi}\frac{1}{n}\sum_{k=1}^{n} m(a, x_k; \phi)\pi_n(\phi)\,\mathrm{d}\phi, \qquad (16)
\end{aligned}$$

where the last form was obtained by choosing the non-parametric specification $\int_{\psi} p(x_i;\psi)\pi_n(\psi)\,\mathrm{d}\psi = p(x_i \mid x_1, \ldots, x_n) = \sum_{k=1}^{n}\delta_{x_k}(x_i)/n$. Alternatively, in (15) one could use the Bayesian bootstrap (Rubin, 1981) specification $p(x_i \mid x_1, \ldots, x_n; \xi) = \sum_{k=1}^{n}\xi_k\delta_{x_k}(x_i)$, where $\xi = (\xi_1, \ldots, \xi_n)$, with $\pi_n(\xi)$ taken to be a uniform Dirichlet distribution (see Section 6). Obtaining (16) also required assuming that $p_{\mathcal{E}}(y_i \mid z_i, x_i; \phi) = p_{\mathcal{O}}(y_i \mid z_i, x_i; \phi) \equiv p(y_i \mid z_i, x_i; \phi)$ and and $p_{\mathcal{E}}(x_i;\psi) = p_{\mathcal{O}}(x_i;\psi) \equiv p(x_i;\psi)$, which corresponds to the stability assumption discussed by Dawid and Didelez (2010).

APPENDIX 2

Proofs to theorems

Proof (to Theorem 1). If (i) holds true, then also $(\mathbf{Y}_{0i}, \mathbf{Y}_{1i}) \perp\!\!\!\perp e(B_i) \mid S_i$, and the propensity score adjustment does not add information. If on the other hand (ii) holds true, $\{e(B_i), S_i\}$ has jointly the balancing property $Z_i \perp\!\!\!\perp X_i \mid \{e(B_i), S_i\}$. This follows from Theorem 2 of Rosenbaum and Rubin (1983, p. 44) and also implies that $(\mathbf{Y}_{0i}, \mathbf{Y}_{1i}) \perp\!\!\!\perp Z_i \mid \{e(B_i), S_i\}$ (Rosenbaum and Rubin, 1983, Theorem 3). Now

$$\begin{aligned}
\mathrm{E}\{Y_i \mid Z_i, e(B_i), S_i\} &= \mathrm{E}\{(1 - Z_i)\mathbf{Y}_{0i} \mid Z_i, e(B_i), S_i\} + \mathrm{E}\{Z_i\mathbf{Y}_{1i} \mid Z_i, e(B_i), S_i\} \\
&= (1 - Z_i)\mathrm{E}\{\mathbf{Y}_{0i} \mid Z_i, e(B_i), S_i\} + Z_i\mathrm{E}\{\mathbf{Y}_{1i} \mid Z_i, e(B_i), S_i\} \\
&= (1 - Z_i)\mathrm{E}\{\mathbf{Y}_{0i} \mid e(B_i), S_i\} + Z_i\mathrm{E}\{\mathbf{Y}_{1i} \mid e(B_i), S_i\},
\end{aligned}$$

and further,

$$\begin{aligned}
&\int_{x_i}\mathrm{E}\{Y_i \mid Z_i = 1, e(b_i), s_i\}p(x_i)\,\mathrm{d}x_i - \int_{x_i}\mathrm{E}\{Y_i \mid Z_i = 0, e(b_i), s_i\}p(x_i)\,\mathrm{d}x_i \qquad (17) \\
&\qquad = \int_{x_i}\mathrm{E}\{\mathbf{Y}_{1i} \mid e(b_i), s_i\}p(x_i)\,\mathrm{d}x_i - \int_{x_i}\mathrm{E}\{\mathbf{Y}_{0i} \mid e(b_i), s_i\}p(x_i)\,\mathrm{d}x_i \\
&\qquad = \mathrm{E}(\mathbf{Y}_{1i}) - \mathrm{E}(\mathbf{Y}_{0i}).
\end{aligned}$$

The consistency of the estimator (3) then relies on being able to consistently estimate the quantities in (17). □

Proof (to Theorem 2). Consider first the expectation of (8) under the assumption that (i) holds true. Now

$$\begin{aligned}
&\mathrm{E}\left\{\mathbf{1}_{\{z_i=a\}}\frac{\frac{\partial}{\partial\phi}\log p(\mathbf{y}_{ai} \mid s_i, \phi)}{\mathrm{pr}(Z_i = a \mid b_i)}\right\} \\
&= \int_{x_i}\int_{\mathbf{y}_{ai}}\sum_{z_i}\mathbf{1}_{\{z_i=a\}}\frac{\frac{\partial}{\partial\phi}\log p(\mathbf{y}_{ai} \mid s_i, \phi)}{\mathrm{pr}(Z_i = a \mid b_i)}f(\mathbf{y}_{ai} \mid z_i, x_i)f(z_i \mid x_i)f(x_i)\,\mathrm{d}\mathbf{y}_{ai}\,\mathrm{d}x_i \\
&= \int_{x_i}\left\{\int_{\mathbf{y}_{ai}}\frac{\frac{\partial}{\partial\phi}p(\mathbf{y}_{ai} \mid s_i; \phi)}{p(\mathbf{y}_{ai} \mid s_i; \phi)}f(\mathbf{y}_{ai} \mid x_i)\,\mathrm{d}\mathbf{y}_{ai}\right\}\frac{\mathrm{pr}(Z_i = a \mid x_i)}{\mathrm{pr}(Z_i = a \mid b_i)}f(x_i)\,\mathrm{d}x_i \\
&= \int_{x_i}\left\{\frac{\partial}{\partial\phi}\int_{\mathbf{y}_{ai}}p(\mathbf{y}_{ai} \mid s_i; \phi)\,\mathrm{d}\mathbf{y}_{ai}\right\}\frac{\mathrm{pr}(Z_i = a \mid x_i)}{\mathrm{pr}(Z_i = a \mid b_i)}f(x_i)\,\mathrm{d}x_i = 0,
\end{aligned}$$

which followed because now $p(\mathbf{y}_{ai} \mid s_i; \phi) = f(\mathbf{y}_{ai} \mid x_i)$ at the true parameter value. Thus, the misspecified weights do not influence the estimation (in terms of bias) as long as the outcome model is correctly specified.

Under the assumption that (ii) holds true, we have in turn that

$$\begin{aligned}
&\mathrm{E}\left\{ \mathbf{1}_{\{z_i=a\}} \frac{\frac{\partial}{\partial\phi} \log p(\mathbf{y}_{ai} \mid s_i; \phi)}{\mathrm{pr}(Z_i = a \mid b_i)} \right\} \\
&= \int_{x_i} \int_{\mathbf{y}_{ai}} \sum_{z_i} \mathbf{1}_{\{z_i=a\}} \frac{\frac{\partial}{\partial\phi} \log p(\mathbf{y}_{ai} \mid s_i; \phi)}{\mathrm{pr}(Z_i = a \mid b_i)} f(\mathbf{y}_{ai} \mid z_i, x_i) f(z_i \mid x_i) f(x_i) \,\mathrm{d}\mathbf{y}_{ai} \,\mathrm{d}x_i \\
&= \int_{x_i} \int_{\mathbf{y}_{ai}} \frac{\partial}{\partial\phi} \log p(\mathbf{y}_{ai} \mid s_i; \phi) f(\mathbf{y}_{ai} \mid x_i) f(x_i) \,\mathrm{d}\mathbf{y}_{ai} \,\mathrm{d}x_i,
\end{aligned}$$

since now $p(z_i \mid b_i) = f(z_i \mid x_i)$. Using the partitioning $x_i = (s_i, r_i)$, we can write the last form in above as

$$\begin{aligned}
&\int_{x_i} \int_{\mathbf{y}_{ai}} \frac{\partial}{\partial\phi} \log p(\mathbf{y}_{ai} \mid s_i; \phi) f(\mathbf{y}_{ai} \mid x_i) f(x_i) \,\mathrm{d}\mathbf{y}_{ai} \,\mathrm{d}x_i \\
&= \int_{s_i} \int_{r_i} \int_{\mathbf{y}_{ai}} \frac{\partial}{\partial\phi} \log p(\mathbf{y}_{ai} \mid s_i; \phi) f(\mathbf{y}_{ai}, s_i, r_i) \,\mathrm{d}\mathbf{y}_{ai} \,\mathrm{d}s_i \,\mathrm{d}r_i \\
&= \int_{s_i} \int_{\mathbf{y}_{ai}} \frac{\frac{\partial}{\partial\phi} p(\mathbf{y}_{ai} \mid s_i; \phi)}{p(\mathbf{y}_{ai} \mid s_i; \phi)} f(\mathbf{y}_{ai} \mid s_i) f(s_i) \,\mathrm{d}\mathbf{y}_{ai} \,\mathrm{d}s_i \\
&= \int_{s_i} \left\{ \frac{\partial}{\partial\phi} \int_{\mathbf{y}_{ai}} p(\mathbf{y}_{ai} \mid s_i; \phi) \,\mathrm{d}\mathbf{y}_{ai} \right\} f(s_i) \,\mathrm{d}s_i = 0.
\end{aligned}$$

Thus, even though the outcome regression does not include a sufficient set of confounders, through the IPT weighting we can still obtain valid estimates for the conditional distributions $p(\mathbf{y}_{ai} \mid s_i; \phi)$. □

Proof (to theorem 3). Now the marginal posterior distribution of the parameters γ becomes

$$\begin{aligned}
p(\gamma \mid v) &= \int_{\phi,\psi} p(\phi, \gamma, \psi \mid v) \,\mathrm{d}\phi \,\mathrm{d}\psi \\
&\propto \int_{\phi,\psi} \left\{ \prod_{i=1}^{n} p(y_i \mid z_i, x_i; \phi) p(z_i \mid x_i; \gamma) p(x_i; \psi) \right\} \pi_0(\phi) \pi_0(\psi) \pi_0(\gamma) \,\mathrm{d}\phi \,\mathrm{d}\psi \,\mathrm{d}\gamma \\
&\propto \prod_{i=1}^{n} p(z_i \mid x_i; \gamma) \pi_0(\gamma) \\
&\propto p(\gamma \mid x, z).
\end{aligned}$$

□

Proof (to Theorem 4). The estimator obtained through substituting in the parametric models is

$$\begin{aligned}
&\sum_{k=1}^{n} \xi_k \left[y_i - m\{z_k, s_k; \widehat{\phi}(\xi)\} \right] \left[\frac{z_k}{\mathrm{pr}_{\mathcal{O}}\{Z_k = 1 \mid b_k; \widehat{\gamma}(\xi)\}} - \frac{1 - z_k}{\mathrm{pr}_{\mathcal{O}}\{Z_k = 0 \mid b_k; \widehat{\gamma}(\xi)\}} \right] \\
&\quad + \sum_{k=1}^{n} \xi_k \left[m\{1, s_k; \widehat{\phi}(\xi)\} - m\{0, s_k; \widehat{\phi}(\xi)\} \right]. \qquad (18)
\end{aligned}$$

First, if the outcome model is correctly specified in the sense that $m\{z_k, s_k; \widehat{\phi}(\xi)\} = m(z_k, x_k; \xi)$, we

get

$$\begin{aligned}
&\mathrm{pr}_{\mathcal{E}}(Z_k = a) \sum_{k=1}^{n} \xi_k \frac{\mathbf{1}_{\{z_k=a\}} m\{z_k, s_k; \widehat{\phi}(\xi)\}}{\mathrm{pr}_{\mathcal{O}}\{Z_k = a \mid b_k; \widehat{\gamma}(\xi)\}} \\
&\quad = \sum_{k=1}^{n} \xi_k \mathbf{1}_{\{z_k=a\}} m\{z_k, s_k; \widehat{\phi}(\xi)\} \frac{p_{\mathcal{E}}(z_k)}{p_{\mathcal{O}}\{z_k \mid b_k; \widehat{\gamma}(\xi)\}} \\
&\quad = \int_{z_i, x_i} \mathbf{1}_{\{z_i=a\}} m\{z_i, s_i; \widehat{\phi}(\xi)\} \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}\{z_i \mid b_i; \widehat{\gamma}(\xi)\}} \sum_{k=1}^{n} \xi_k \delta_{(z_k, x_k)}(z_i, x_i) \, \mathrm{d}z_i \, \mathrm{d}x_i \\
&\quad = \int_{y_i, z_i, x_i} \mathbf{1}_{\{z_i=a\}} y_i \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}\{z_i \mid b_i; \widehat{\gamma}(\xi)\}} p_{\mathcal{O}}(y_i \mid z_i, x_i, v; \xi) p_{\mathcal{O}}(z_i, x_i \mid v; \xi) \, \mathrm{d}y_i \, \mathrm{d}z_i \, \mathrm{d}x_i \\
&\quad = \int_{v_i} \mathbf{1}_{\{z_i=a\}} y_i \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}\{z_i \mid b_i; \widehat{\gamma}(\xi)\}} p_{\mathcal{O}}(v_i \mid v; \xi) \, \mathrm{d}v_i \\
&\quad = \mathrm{pr}_{\mathcal{E}}(Z_k = a) \sum_{k=1}^{n} \xi_k \frac{\mathbf{1}_{\{z_k=a\}} y_k}{\mathrm{pr}_{\mathcal{O}}\{Z_k = a \mid b_k; \widehat{\gamma}(\xi)\}},
\end{aligned}$$

because the second to last form is equivalent to (19) in Supplementary Appendix 4. Thus, the first summation term in (18) cancels out, leaving only the model based estimator, which itself is now equivalent to the posterior predictive mean difference.

On the other hand, if the treatment assignment model is correctly specified in the sense that $\mathrm{pr}_{\mathcal{O}}\{Z_k = a \mid b_k; \widehat{\gamma}(\xi)\} = \mathrm{pr}_{\mathcal{O}}(Z_k = a \mid x_k, v; \xi)$, we get

$$\begin{aligned}
&\mathrm{pr}_{\mathcal{E}}(Z_k = a) \sum_{k=1}^{n} \xi_k \frac{\mathbf{1}_{\{z_k=a\}} m\{z_k, s_k; \widehat{\phi}(\xi)\}}{\mathrm{pr}_{\mathcal{O}}\{Z_k = a \mid b_k; \widehat{\gamma}(\xi)\}} \\
&\quad = \sum_{k=1}^{n} \xi_k \mathbf{1}_{\{z_k=a\}} m\{z_k, s_k; \widehat{\phi}(\xi)\} \frac{p_{\mathcal{E}}(z_k)}{p_{\mathcal{O}}\{z_k \mid b_k; \widehat{\gamma}(\xi)\}} \\
&\quad = \int_{z_i, x_i} \mathbf{1}_{\{z_i=a\}} m\{z_i, s_i; \widehat{\phi}(\xi)\} \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}\{z_i \mid b_i; \widehat{\gamma}(\xi)\}} \sum_{k=1}^{n} \xi_k \delta_{(z_k, x_k)}(z_i, x_i) \, \mathrm{d}z_i \, \mathrm{d}x_i \\
&\quad = \int_{z_i, x_i} \mathbf{1}_{\{z_i=a\}} m\{z_i, s_i; \widehat{\phi}(\xi)\} \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}\{z_i \mid b_i; \widehat{\gamma}(\xi)\}} p_{\mathcal{O}}(z_i \mid x_i, v; \xi) p_{\mathcal{O}}(x_i \mid v; \xi) \, \mathrm{d}z_i \, \mathrm{d}x_i \\
&\quad = \int_{z_i, x_i} \mathbf{1}_{\{z_i=a\}} m\{z_i, s_i; \widehat{\phi}(\xi)\} p_{\mathcal{E}}(z_i) p_{\mathcal{O}}(x_i \mid v; \xi) \, \mathrm{d}z_i \, \mathrm{d}x_i \\
&\quad = \mathrm{pr}_{\mathcal{E}}(Z_i = a) \int_{x_i} m\{a, s_i; \widehat{\phi}(\xi)\} \sum_{k=1}^{n} \xi_k \delta_{(x_k)}(x_i) \, \mathrm{d}x_i \\
&\quad = \mathrm{pr}_{\mathcal{E}}(Z_k = a) \sum_{k=1}^{n} \xi_k m\{a, s_k; \widehat{\phi}(\xi)\}.
\end{aligned}$$

Therefore, the estimator (18) now reduces to

$$\sum_{k=1}^{n} \xi_k y_i \left[\frac{z_k}{\mathrm{pr}_{\mathcal{O}}\{Z_k = 1 \mid b_k; \widehat{\gamma}(\xi)\}} - \frac{1 - z_k}{\mathrm{pr}_{\mathcal{O}}\{Z_k = 0 \mid b_k; \widehat{\gamma}(\xi)\}} \right],$$

which is again equivalent to the posterior predictive mean difference (see Supplementary Appendix 4).□

APPENDIX 3

On the frequency-based properties of the two-step approach

Trivially, if the outcome model is correctly specified, then $\Phi \perp\!\!\!\perp \Gamma \mid V$ and (9) reduces to (16). The interesting situations are naturally those where this is not the case. We denote the log-likelihood by $q_i(\phi; \gamma) = \log p\{y_i \mid z_i, e(b_i; \gamma), s_i; \phi)\}$ and $q(\phi; \gamma) \equiv \sum_{i=1}^{n} q_i(\phi; \gamma)$ and consider the quasi-maximum likelihood estimator $\widehat{\phi}(\widehat{\gamma}) \equiv \arg\max_{\phi} q(\phi; \widehat{\gamma})$. If the treatment assignment model is correctly specified,

$\widehat{\gamma} \to \gamma_0$. In addition, we assume that with any fixed value of γ, $\widehat{\phi}(\gamma) \to \phi_0(\gamma)$, where $\phi_0(\gamma)$ is the parameter vector which minimizes the Kullback-Leibler divergence to the true outcome model (e.g. White, 1982, p. 4). Thus, by the law of large numbers and continuous mapping, we can write in the usual way that

$$\frac{1}{n}\sum_{i=1}^{n}[q_i(\phi;\widehat{\gamma}) - q_i\{\phi_0(\gamma_0);\gamma_0\}] \to E\left[q_i(\phi;\gamma_0) - q_i\{\phi_0(\gamma_0);\gamma_0\}\right],$$

where the right hand side is maximized at zero when $\phi = \phi_0(\gamma_0)$, at which point

$$\mathrm{E}\{Y_i \,|\, Z_i = a, e(b_i;\gamma_0), s_i; \phi_0(\gamma_0)\} = \mathrm{E}\{\mathbf{Y}_{ia} \,|\, e(b_i;\gamma_0), s_i; \phi_0(\gamma_0)\}.$$

Since we also have that the posterior $p(\gamma \,|\, x, y) \to \delta_{\gamma_0}(\gamma)$ in distribution, we can then conjecture that posterior predictive inferences based on (9) will be asymptotically uncounfounded.

With the definitions

$$\begin{aligned}
U^{\phi}(\phi;\gamma) &\equiv \partial q(\phi;\gamma)/\partial\phi,\\
U^{\phi\phi}(\phi;\gamma) &\equiv \partial^2 q(\phi;\gamma)/\partial\phi^2,\\
U^{\phi\gamma}(\phi;\gamma) &\equiv \partial^2 q(\phi;\gamma)/\partial\phi\partial\gamma,\\
U^{\gamma}(\gamma) &\equiv \partial \textstyle\sum_{i=1}^{n}\log p(z_i \,|\, b_i;\gamma)/\partial\gamma,\\
U^{\gamma\gamma}(\gamma) &\equiv \partial^2 \textstyle\sum_{i=1}^{n}\log p(z_i \,|\, b_i;\gamma)/\partial\gamma^2,
\end{aligned}$$

and noting that $U^{\phi}(\widehat{\phi};\gamma^{(j)}) = 0$ for each $\gamma^{(j)}$, $j = 1,\ldots,m$, we can consider the first order Taylor expansion of $U^{\phi}(\widehat{\phi};\gamma^{(j)})$ around the true parameter values (ϕ_0,γ_0), which becomes

$$\begin{aligned}
0 &= \frac{1}{n}U^{\phi}(\widehat{\phi};\gamma^{(j)})\\
&\approx \frac{1}{n}U^{\phi}(\phi_0;\gamma_0) + \frac{1}{n}U^{\phi\phi}(\phi_0;\gamma_0)\{\widehat{\phi}(\gamma^{(j)}) - \phi_0\} + \frac{1}{n}U^{\phi\gamma}(\phi_0;\gamma_0)(\gamma^{(j)} - \gamma_0)\\
&\approx \frac{1}{n}U^{\phi}(\phi_0;\gamma_0) + \mathrm{E}\{U_i^{\phi\phi}(\phi_0;\gamma_0)\}\{\widehat{\phi}(\gamma^{(j)}) - \phi_0\} + \mathrm{E}\{U_i^{\phi\gamma}(\phi_0;\gamma_0)\}(\gamma^{(j)} - \gamma_0),
\end{aligned}$$

and further,

$$\begin{aligned}
0 &= \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}U^{\phi}(\widehat{\phi};\gamma^{(j)})\\
&\approx \frac{1}{n}U^{\phi}(\phi_0;\gamma_0) + \mathrm{E}\{U_i^{\phi\phi}(\phi_0;\gamma_0)\}\left\{\frac{1}{m}\sum_{j=1}^{m}\widehat{\phi}(\gamma^{(j)}) - \phi_0\right\} + \mathrm{E}\{U_i^{\phi\gamma}(\phi_0;\gamma_0)\}\left(\widehat{\gamma} - \gamma_0\right),
\end{aligned}$$

if $\frac{1}{m}\sum_{j=1}^{m}\gamma^{(j)} \approx \widehat{\gamma}$. Hence,

$$\begin{aligned}
\sqrt{n}\left\{\frac{1}{m}\sum_{j=1}^{m}\widehat{\phi}(\gamma^{(j)}) - \phi_0\right\} &\approx \mathrm{E}\{-U_i^{\phi\phi}(\phi_0;\gamma_0)\}^{-1}\\
&\quad\times\left[\frac{\sqrt{n}}{n}U^{\phi}(\phi_0;\gamma_0) + \mathrm{E}\{U_i^{\phi\gamma}(\phi_0;\gamma_0)\}\sqrt{n}(\widehat{\gamma} - \gamma_0)\right].
\end{aligned}$$

Here we have, by another first order expansion around γ_0, that

$$\sqrt{n}(\widehat{\gamma} - \gamma_0) \approx \mathrm{E}\{-U_i^{\gamma\gamma}(\gamma_0)\}^{-1}\frac{\sqrt{n}}{n}U^{\gamma}(\gamma_0),$$

so finally,

$$\begin{aligned}
\sqrt{n}\left\{\frac{1}{m}\sum_{j=1}^{m}\widehat{\phi}(\gamma^{(j)}) - \phi_0\right\} &\approx \mathrm{E}\{-U_i^{\phi\phi}(\phi_0;\gamma_0)\}^{-1}\\
&\quad\times\left(\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left[U_i^{\phi}(\phi_0;\gamma_0) + \mathrm{E}\{U_i^{\phi\gamma}(\phi_0;\gamma_0)\}\mathrm{E}\{-U_i^{\gamma\gamma}(\gamma_0)\}^{-1}U_i^{\gamma}(\gamma_0)\right]\right).
\end{aligned}$$

We may similarly expand $U^{\phi}(\widehat{\phi};\widehat{\gamma})$ where the parameters γ have been fixed to their maximum likelihood estimates around (ϕ_0,γ_0) as

$$\begin{aligned}0 &= \frac{1}{n}U^{\phi}(\widehat{\phi};\widehat{\gamma}) \\ &\approx \frac{1}{n}U^{\phi}(\phi_0;\gamma_0) + \mathrm{E}\{U_i^{\phi\phi}(\phi_0;\gamma_0)\}(\widehat{\phi}(\widehat{\gamma})-\phi_0) + \mathrm{E}\{U_i^{\phi\gamma}(\phi_0;\gamma_0)\}(\widehat{\gamma}-\gamma_0)\end{aligned}$$

to find that

$$\sqrt{n}\left\{\widehat{\phi}(\widehat{\gamma})-\phi_0\right\} \approx \mathrm{E}\{-U_i^{\phi\phi}(\phi_0;\gamma_0)\}^{-1}\frac{\sqrt{n}}{n}\sum_{i=1}^{n}B_i(\phi_0;\gamma_0),$$

where

$$B_i(\phi_0;\gamma_0) \equiv U_i^{\phi}(\phi_0;\gamma_0) + \mathrm{E}\{U_i^{\phi\gamma}(\phi_0;\gamma_0)\}\mathrm{E}\{-U_i^{\gamma\gamma}(\gamma_0)\}^{-1}U_i^{\gamma}(\gamma_0).$$

Since the two estimators $\frac{1}{m}\sum_{j=1}^{m}\widehat{\phi}(\gamma^{(j)})$ and $\widehat{\phi}(\widehat{\gamma})$ have the same linear approximation which is a sum of independent terms, we conclude that they have the same asymptotic distribution,

$$\sqrt{n}(\widehat{\phi}-\phi_0) \to \mathrm{N}\left[0, \mathrm{E}\{-U_i^{\phi\phi}(\phi_0;\gamma_0)\}^{-1}\mathrm{var}\{B_i(\phi_0;\gamma_0)\}\mathrm{E}\{-U^{\phi\phi}(\phi_0;\gamma_0)^{\top}\}^{-1}\right].$$

Fitting the regression model $y_i = \phi_0 + \phi_1 z_i + \phi_2^{\top}s_i + \phi_3^{\top}g\{e(b_i;\gamma)\} + \varepsilon_{1i}$ to estimate the parameter of interest ϕ_1 is numerically equivalent to fitting the sequence of regressions $y_i = \nu^{\top}s_i^*(\widehat{\gamma}) + \varepsilon_{2i}$, $z_i = \alpha^{\top}s_i^*(\widehat{\gamma}) + \varepsilon_{3i}$ and $\{y_i - \widehat{\nu}^{\top}s_i^*(\widehat{\gamma})\} = \phi_1\{z_i - \widehat{\alpha}^{\top}s_i^*(\widehat{\gamma})\} + \varepsilon_{4i}$, where $s_i^*(\gamma) \equiv [s_i, g\{e(b_i;\gamma)\}]$. Denoting the estimating function corresponding to the last regression as

$$U^{\phi_1}\{\phi_1,\widehat{\gamma},\widehat{\nu}(\widehat{\gamma}),\widehat{\alpha}(\widehat{\gamma})\} \equiv \sum_{i=1}^{n}\{z_i - \widehat{\alpha}^{\top}s_i^*(\widehat{\gamma})\}\left[\{y_i - \widehat{\nu}^{\top}s_i^*(\widehat{\gamma})\} - \phi_1\{z_i - \widehat{\alpha}^{\top}s_i^*(\widehat{\gamma})\}\right],$$

and the partial derivatives of this as e.g. $U^{\phi_1\gamma} \equiv \partial U^{\phi_1}/\partial\gamma$, we can expand this around $(\phi_{10},\gamma_0,\nu_0,\alpha_0)$, where $\nu_0 \equiv \nu_0(\gamma_0)$ and $\alpha_0 \equiv \alpha_0(\gamma_0)$ are the limiting values of the nuisance parameters, as

$$\begin{aligned}\frac{1}{n}U^{\phi_1}\{\phi_1,\widehat{\gamma},\widehat{\nu}(\widehat{\gamma}),\widehat{\alpha}(\widehat{\gamma})\} &\approx \frac{1}{n}U^{\phi_1}(\phi_{10},\gamma_0,\nu_0,\alpha_0) + \mathrm{E}\{U_i^{\phi_1\gamma}(\phi_{10},\gamma_0,\nu_0,\alpha_0)\}(\widehat{\gamma}-\gamma_0) \\ &\quad + \mathrm{E}\{U_i^{\phi_1\nu}(\phi_{10},\gamma_0,\nu_0,\alpha_0)\}(\widehat{\nu}-\gamma_0) \\ &\quad + \mathrm{E}\{U_i^{\phi_1\alpha}(\phi_{10},\gamma_0,\nu_0,\alpha_0)\}(\widehat{\alpha}-\gamma_0) \\ &= \frac{1}{n}U^{\phi_1}(\phi_{10},\gamma_0,\nu_0,\alpha_0) + \mathrm{E}\{U_i^{\phi_1\gamma}(\phi_{10},\gamma_0,\nu_0,\alpha_0)\}(\widehat{\gamma}-\gamma_0) \\ &\approx \frac{1}{n}U^{\phi_1}(\phi_1,\widehat{\gamma},\nu_0,\alpha_0),\end{aligned}$$

since here $\mathrm{E}\{U_i^{\phi_1\nu}(\phi_{10},\gamma_0,\nu_0,\alpha_0)\} = \mathrm{E}\{U_i^{\phi_1\alpha}(\phi_{10},\gamma_0,\nu_0,\alpha_0)\} = 0$. We can now see that the Theorem 1 of Henmi and Eguchi (2004, p. 935) applies to the last form here, implying that $\mathrm{avar}(\widehat{\phi}_1) \leq \mathrm{avar}(\tilde{\phi}_1)$, where $\widehat{\phi}_1$ is the solution to $U(\phi_1,\widehat{\gamma},\nu_0,\alpha_0)=0$ and $\tilde{\phi}_1$ is the solution to $U(\phi_1,\gamma_0,\nu_0,\alpha_0)=0$.

APPENDIX 4

The doubly robust estimator as a posterior predictive expectation

We first note that because

$$\begin{aligned}&\int_{v_i}\mathbf{1}_{\{z_i=a\}}y_i p_{\mathcal{E}}(v_i \mid v;\xi)\,\mathrm{d}v_i \\ &\quad= \int_{y_i,x_i}y_i p_{\mathcal{E}}(y_i \mid Z_i=a,x_i,v;\xi)\mathrm{pr}_{\mathcal{E}}(Z_i=a)p_{\mathcal{E}}(x_i \mid v;\xi)\,\mathrm{d}y_i\,\mathrm{d}x_i,\end{aligned}$$

we have that

$$\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i=1,v;\xi) = \mathrm{E}_{\mathcal{E}}\left\{\frac{Z_iY_i}{\mathrm{pr}_{\mathcal{E}}(Z_i=1)}\,\middle|\, v;\xi\right\}$$

and

$$\mathrm{E}_{\mathcal{E}}(Y_i \,|\, Z_i = 0, v; \xi) = \mathrm{E}_{\mathcal{E}} \left\{ \frac{(1 - Z_i)Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 0)} \,\middle|\, v; \xi \right\}.$$

The usual IPT-weighted estimator for a marginal causal contrast may be derived through a posterior predictive argument as follows. First,

$$\begin{aligned}
\mathrm{E}_{\mathcal{E}}[Z_i Y_i \,|\, v; \xi] &= \int_{v_i} z_i y_i p_{\mathcal{E}}(v_i \,|\, v; \xi) \,\mathrm{d}v_i \\
&= \int_{v_i} z_i y_i \frac{p_{\mathcal{E}}(v_i \,|\, v; \xi)}{p_{\mathcal{O}}(v_i \,|\, v; \xi)} p_{\mathcal{O}}(v_i \,|\, v; \xi) \,\mathrm{d}v_i \\
&= \int_{v_i} z_i y_i \frac{p_{\mathcal{E}}(y_i \,|\, z_i, x_i, v; \xi) p_{\mathcal{E}}(z_i) p_{\mathcal{E}}(x_i \,|\, v, \xi)}{p_{\mathcal{O}}(y_i \,|\, z_i, x_i, v; \xi) p_{\mathcal{O}}(z_i \,|\, x_i, v; \xi) p_{\mathcal{O}}(x_i \,|\, v; \xi)} p_{\mathcal{O}}(v_i \,|\, v; \xi) \,\mathrm{d}v_i \\
&= \int_{v_i} z_i y_i \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}(z_i \,|\, x_i, v; \xi)} p_{\mathcal{O}}(v_i \,|\, v; \xi) \,\mathrm{d}v_i \qquad (19) \\
&= \int_{v_i} z_i y_i \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}(z_i \,|\, x_i, v; \xi)} \sum_{k=1}^{n} \xi_k \delta_{v_k}(v_i) \,\mathrm{d}v_i \\
&= \sum_{k=1}^{n} \xi_k z_k y_k \frac{p_{\mathcal{E}}(z_k)}{p_{\mathcal{O}}(z_k \,|\, x_k, v; \xi)} \\
&= \mathrm{pr}_{\mathcal{E}}(Z_k = 1) \sum_{k=1}^{n} \xi_k \frac{z_k y_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \,|\, x_k, v; \xi)},
\end{aligned}$$

and thus,

$$\mathrm{E}_{\mathcal{E}} \left\{ \frac{Z_i Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 1)} \,\middle|\, v; \xi \right\} = \sum_{k=1}^{n} \xi_k \frac{z_k y_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \,|\, x_k, v; \xi)}.$$

Similarly,

$$\mathrm{E}_{\mathcal{E}} \left\{ \frac{(1 - Z_i)Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 0)} \,\middle|\, v; \xi \right\} = \sum_{k=1}^{n} \xi_k \frac{(1 - z_k) y_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 0 \,|\, x_k, v; \xi)},$$

and

$$\begin{aligned}
&\mathrm{E}_{\mathcal{E}} \left\{ \frac{Z_i Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 1)} - \frac{(1 - Z_i)Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 0)} \,\middle|\, v; \xi \right\} \\
&\quad = \sum_{k=1}^{n} \xi_k y_k \left\{ \frac{z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \,|\, x_k, v; \xi)} - \frac{1 - z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 0 \,|\, x_k, v; \xi)} \right\}.
\end{aligned}$$

On the other hand, the usual outcome model based estimator may be motivated similarly as in Supplementary Appendix 1 through

$$\begin{aligned}
\mathrm{E}_{\mathcal{E}}(Y_i \,|\, Z_i = a, v; \xi) &= \int_{x_i} \left\{ \int_{y_i} y_i p_{\mathcal{O}}(y_i \,|\, Z_i = a, x_i, v; \xi) \,\mathrm{d}y_i \right\} p_{\mathcal{O}}(x_i \,|\, v; \xi) \,\mathrm{d}x_i \\
&= \int_{x_i} m(a, x_i; \xi) \sum_{k=1}^{n} \xi_k \delta_{x_k}(x_i) \,\mathrm{d}x_i \\
&= \sum_{k=1}^{n} \xi_k m(a, x_k; \xi),
\end{aligned}$$

and

$$\mathrm{E}_{\mathcal{E}}(Y_i \,|\, Z_i = 1, v; \xi) - \mathrm{E}_{\mathcal{E}}(Y_i \,|\, Z_i = 0, v; \xi) = \sum_{k=1}^{n} \xi_k \left\{ m(1, x_k; \xi) - m(0, x_k; \xi) \right\}.$$

Finally, we note that we can write (19) alternatively as

$$
\begin{aligned}
&\mathrm{E}_{\mathcal{E}}(Z_i Y_i \mid v;\xi) \\
&\quad= \int_{v_i} z_i y_i \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}(z_i \mid x_i, v;\xi)} p_{\mathcal{O}}(v_i \mid v;\xi)\,\mathrm{d}v_i \\
&\quad= \int_{y_i,z_i,x_i} z_i y_i \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}(z_i \mid x_i, v;\xi)} p_{\mathcal{O}}(y_i \mid z_i, x_i, v;\xi) p_{\mathcal{O}}(z_i, x_i \mid v;\xi)\,\mathrm{d}y_i\,\mathrm{d}z_i\,\mathrm{d}x_i \\
&\quad= \int_{z_i,x_i} z_i m(z_i, x_i;\xi) \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}(z_i \mid x_i, v;\xi)} \sum_{k=1}^{n} \xi_k \delta_{(z_k,x_k)}(z_i, x_i)\,\mathrm{d}z_i\,\mathrm{d}x_i \\
&\quad= \sum_{k=1}^{n} \xi_k z_k m(z_k, x_k;\xi) \frac{p_{\mathcal{E}}(z_k)}{p_{\mathcal{O}}(z_i \mid x_k, v;\xi)} \\
&\quad= \mathrm{pr}_{\mathcal{E}}(Z_k = 1) \sum_{k=1}^{n} \xi_k \frac{z_k m(z_k, x_k;\xi)}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \mid x_k, v;\xi)},
\end{aligned}
$$

and therefore

$$
\begin{aligned}
&\mathrm{E}_{\mathcal{E}}\left\{ \frac{Z_i Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 1)} - \frac{(1 - Z_i) Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 0)} \;\middle|\; v;\xi \right\} \\
&\qquad= \sum_{k=1}^{n} \xi_k m(z_k, x_k;\xi) \left\{ \frac{z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \mid x_k, v;\xi)} - \frac{1 - z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 0 \mid x_k, v;\xi)} \right\}.
\end{aligned}
$$

Thus, the posterior predictive mean difference can be written as

$$
\begin{aligned}
&\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 1, v;\xi) - \mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 0, v;\xi) \\
&\quad= \mathrm{E}_{\mathcal{E}}\left\{ \frac{Z_i Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 1)} - \frac{(1 - Z_i) Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 0)} \;\middle|\; v;\xi \right\} + \mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 1, v;\xi) \\
&\qquad- \mathrm{E}_{\mathcal{E}}\left\{ \frac{Z_i Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 1)} - \frac{(1 - Z_i) Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 0)} \;\middle|\; v;\xi \right\} - \mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 0, v;\xi) \\
&\quad= \sum_{k=1}^{n} \xi_k \{y_i - m(z_k, x_k;\xi)\} \left\{ \frac{z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \mid x_k, v;\xi)} - \frac{1 - z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 0 \mid x_k, v;\xi)} \right\} \\
&\qquad+ \sum_{k=1}^{n} \xi_k \left\{ m(1, x_k;\xi) - m(0, x_k;\xi) \right\}. \qquad (20)
\end{aligned}
$$

REFERENCES

Achy-Brou, A. C., Frangakis, C. E., and Griswold, M. (2010). Estimating treatment effects of longitudinal designs using regression models on propensity scores. *Biometrics*, 66:824–833.

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

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, Chichester.

Chakraborty, B. and Moodie, E. E. M. (2013). *Statistical Methods for Dynamic Treatment Regimes*. Springer-Verlag, New York.

Chamberlain, G. and Imbens, G. W. (2003). Nonparametric applications of Bayesian inference. *Journal of Business & Economic Statistics*, 21:12–18.

Chen, J. and Kaplan, D. (2015). Covariate balance in Bayesian propensity score approaches for observational studies. *Journal of Research on Educational Effectiveness*, 8:280–302.

Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168:656–664.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B*, 41:1–31.

Dawid, A. P. and Didelez, V. (2010). Identifying the consequences of dynamic treatment strategies: a decision-theoretic overview. *Statistical Surveys*, 4:184–231.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7:1–26.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, Boca Raton, FL.

Graham, D. J., McCoy, E. J., and Stephens, D. A. (2015). Approximate Bayesian inference for doubly robust estimation. *Bayesian Analysis*. doi:10.1214/14-BA928.

Gustafson, P. (2012). Double-robust estimators: slightly more Bayesian than meets the eye? *The International Journal of Biostatistics*, 8. DOI: 10.2202/1557-4679.1349.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95:481–488.

Henmi, M. and Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika*, 91:929–941.

Hernán, M. A., Brumback, B., and Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96:440–448.

Hernán, M. A. and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60:578–586.

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

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17:295–311.

Kaplan, D. and Chen, J. (2012). Two-step bayesian approach for propensity score analysis: simulations and case study. *Psychometrika*, 77:581–609.

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., and Neuenschwander, B. (2009). Combining MCMC with "sequential" PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics*, 36:19–39.

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

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

McCandless, L. C., Gustafson, P., Austin, P. C., and Levy, A. R. (2009b). Covariate balance in a Bayesian propensity score analysis of beta blocker therapy in heart failure patients. *Epidemiologic Perspectives & Innovations*, 6.

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

Newton, M. A. and Raftery, A. E. (1994). Approximating Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B*, 56:3–48.

Robins, J. M., Hernán, M. Á., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16:285–319.

Rose, S. and van der Laan, M. J. (2008). Simple optimal weighting of cases and controls in case-control studies. *The International Journal of Biostatistics*, 4. DOI: 10.2202/1557-4679.1115.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 6:41–55.

Røysland, K. (2011). A martingale approach to continuous-time marginal structural models. *Bernoulli*, 17:895–915.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9:130–134.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489.

Saarela, O., Arjas, E., Stephens, D. A., and Moodie, E. E. M. (2015a). Predictive Bayesian inference and dynamic treatment regimes. *Biometrical Journal*. 10.1002/bimj.201400153.

Saarela, O., Moodie, E. E. M., Stephens, D. A., and Klein, M. B. (2015b). On Bayesian estimation of marginal structural models. *Biometrics*. doi:10.1111/biom.12269.

Saarela, O., Moodie, E. E. M., Stephens, D. A., and Klein, M. B. (2015c). Rejoinder: On Bayesian estimation of marginal structural models. *Biometrics*. doi:10.1111/biom.12274.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94:1096–1146.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York, NY.

Walker, S. G. (2010). Bayesian nonparametric methods: motivation and ideas. In Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., editors, *Bayesian Nonparametrics*. Cambridge University Press, Cambridge, UK.

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

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 115:1–25.

Xiao, Y., Moodie, E. E. M., and Abrahamowicz, M. (2013). Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods*, 2. doi:10.1515/em-2012-0006.

Zhang, G. and Little, R. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65:911–918.

Zigler, C. M. and Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, 109(505):95–107.

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.