DOI: 10.1111/biom.12269

On Bayesian Estimation of Marginal Structural Models

Olli Saarela,[1],* David A. Stephens,[2] Erica E. M. Moodie,[3] and Marina B. Klein[4]

[1]Dalla Lana School of Public Health, University of Toronto, 155 College Street, 6th floor, Toronto, Ontario, Canada M5T 3M7

[2]Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montreal, Quebec, Canada H3A 2K6

[3]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, Quebec, Canada H3A 1A2

[4]Department of Medicine, Division of Infectious Diseases, McGill University, 3650 Saint Urbain, Montreal, Quebec, Canada H2X 2P4

**email:* olli.saarela@utoronto.ca

Summary. The purpose of inverse probability of treatment (IPT) weighting in estimation of marginal treatment effects is to construct a pseudo-population without imbalances in measured covariates, thus removing the effects of confounding and informative censoring when performing inference. In this article, we formalize the notion of such a pseudo-population as a data generating mechanism with particular characteristics, and show that this leads to a natural Bayesian interpretation of IPT weighted estimation. Using this interpretation, we are able to propose the first fully Bayesian procedure for estimating parameters of marginal structural models using an IPT weighting. Our approach suggests that the weights should be derived from the posterior predictive treatment assignment and censoring probabilities, answering the question of whether and how the uncertainty in the estimation of the weights should be incorporated in Bayesian inference of marginal treatment effects. The proposed approach is compared to existing methods in simulated data, and applied to an analysis of the Canadian Co-infection Cohort.

Key words: Bayesian inference; Causal inference; Inverse probability weighting; Longitudinal data; Marginal structural models; Posterior predictive inference; Variance estimation.

1. Introduction

Propensity score adjustment (Rosenbaum and Rubin, 1983), in the form of either weighting, matching, stratification, or covariate adjustment, provides a way to control for confounding in non-experimental settings without having to model the dependence between the confounders and the outcome of interest, given that the probability of the treatment assignment can be correctly modeled with respect to confounding variables. Adjustment via the propensity score is typically carried out in a two stage procedure: first, a parametric propensity score model for the treatment given the covariates is proposed, and parameters estimated from the observed data; second, appropriately comparable individuals—so assessed using the estimated propensity score—are compared in order to assess the unconfounded effect of treatment. The two stage estimation is most commonly justified, and studied theoretically, using frequentist semiparametric theory. It is not typically regarded as being derived from a likelihood-based paradigm.

Bayesian inference, on the other hand, **always** derives from a full probability model specification, which is why, in general, propensity score adjustment methods do not appear to have obvious Bayesian counterparts. For matching methods, there is no clearly defined joint probability model for the observable quantities; for covariate adjustment using the propensity score (or outcome regression) the presumed likelihood is based on a patently misspecified model, as the propensity score predictor cannot readily be thought of as a genuine component of the data generating process. For inverse weighting-based adjustments, no fully Bayesian justification has yet been proposed; we aim to fill this gap in the literature.

The recent causal inference literature has seen several attempts to introduce Bayesian versions of propensity score based methods, including inverse probability of treatment (IPT) weighting (Hoshino, 2008; Kaplan and Chen, 2012), covariate adjustment (McCandless, Gustafson, and Austin, 2009; McCandless et al., 2010; Zigler et al., 2013) and matching (An, 2010). In this article, we provide a fully Bayesian argument that gives further insight into aspects of the previously proposed approaches. Our specific focus will be on IPT weighting in the context of *marginal structural models* (MSMs, Robins, Hernán, and Brumback, 2000; Hernán, Brumback, and Robins, 2001).

The advantage of a marginal model specification, coupled with weighting, is that in addition to controlling for measured confounding, due to the marginalization over the covariate distribution, the impact of any related mediation and effect modification need not be modeled explicitly. Under longitudinal settings, explicit modeling and integration over the (possibly high dimensional) intermediate variables represents a formidable task even in simple settings, and this is why the ability to circumvent this modeling step appears to be an important advantage that inverse probability weighted methods have over Bayesian inferences based on fully specified probability models.

Our motivating example is introduced in Section 2, while in Section 3 we propose a Bayesian interpretation of IPT weighting and a corresponding estimation approach. Since IPT weighted estimation can be interpreted as construction of a pseudo-population with measured covariate imbalances removed, a Bayesian version of the procedure can be linked to sampling from such a pseudo-population, and a Bayes decision rule derived from a change of probability measure, or equivalently, an importance sampling argument. The resulting inference procedure is related to the relevance weighted likelihood of Hu and Zidek (2002) and Wang (2006) and the weighted likelihood bootstrap of Newton and Raftery (1994).

We contrast the fully Bayesian procedure to some existing Bayesian proposals in Section 4. It is a well-known (e.g., Hernán et al., 2001; Henmi and Eguchi, 2004) result that an IPT weighted estimator with estimated weights has a smaller asymptotic variance than the corresponding estimator with the true weights known, which can be intuitively understood in terms of the sample balance given by the estimated propensity score (Rosenbaum and Rubin, 1983, p. 47). However, many of the approaches suggested for Bayesian propensity score adjustment (e.g., Kaplan and Chen, 2012, p. 592) incorporate an additional variance component acknowledging the estimation of the propensity scores. We identify the source of this apparent anomaly to be the lack of a well defined joint probability distribution. In Section 5, we investigate the frequency-based properties of the different Bayesian approaches in a simulation study. In Section 6, we analyze data from the Canadian HIV/Hepatitis C Co-Infection Cohort Study. We conclude with a discussion in Section 7.

2. Motivating Example: Antiretroviral Therepy Interruption and Liver Fibrosis in HIV/HCV Co-Infected Individuals

Our motivating example is a complex longitudinal data set relating to health outcomes for individuals simultaneously infected with HIV and the hepatitis C virus (HCV), in particular, the possible negative influence of treatment interruption on specific endpoints. Although antiretroviral therapy (ART) has reduced morbidity and mortality due to nearly all HIV-related illnesses, this is not the case for mortality due to end-stage liver disease, which has increased since ART treatment became widespread (Klein et al., 2010, p. 1162). In part, this increase may be due to improved overall survival combined with HCV associated hepatic liver fibrosis, the progress of which is accelerated by immune dysfunction related to HIV-infection. The Canadian Co-infection Cohort (CCC) Study (Klein et al., 2010) is one of the largest projects set up to study the role of ART on the development of end-stage liver disease in HIV–HCV co-infected individuals. Given the importance of ART in improving HIV-related immunosuppression, it is hypothesized (Thorpe et al., 2011, p. 968) that liver fibrosis progression in co-infected individuals may be partly related to adverse consequences of ART interruptions. The available data constitute health information for over a thousand co-infected individuals recorded longitudinally over a series of clinic visits, which take place at approximately 6-month intervals.

The objective of our analysis is to assess the causal effect of ART interruption in a between-clinic visit interval on progression to liver fibrosis. As in the majority of observational data sets, there is a strong suggestion of possible confounding, in that factors that influence ART interruption in any interval—for example, involvement in risky lifestyle practices such as intravenous drug use or alcohol abuse—also are likely to induce liver fibrosis. Furthermore, the effect ART interruption in one interval may be felt directly but also be mediated through subsequent health status, and also it may influence subsequent ART interruption incidents.

In the presence of both time-varying confounding and mediation, estimation of the (marginal) causal effect of interest via standard regression methods is not possible, motivating marginal structural modeling. However, from a Bayesian perspective, such procedures seem potentially problematic, as there is no corresponding likelihood function. Our methodological objective in this article is to provide a formal Bayesian justification and estimation procedure for MSMs.

3. A Bayesian Formulation and Interpretation of IPT Weighting

3.1. *Marginal Structural Models*

Consider a longitudinal observational study setting involving the individuals $i = 1, \dots, n$, with measurements of covariates and subsequent treatment decisions carried out at discrete time points $j = 1, \dots, m$. Let $\tilde{z}_i \equiv (z_{i1}, z_{i2}, \dots, z_{im})$ denote the observed history of treatment assignments or prescribed doses. Further, let y_i be the outcome of interest observed after sufficient time has passed from the last time-point, and $\tilde{x}_i \equiv (x_{i1}, x_{i2}, \dots, x_{im})$ denote an observed history of vectors of covariates, including a sufficient set of (possibly time-dependent) confounders, recorded before each treatment assignment. Partial histories up to and including timepoint j are denoted as, for example, $\tilde{x}_{ij} \equiv (x_{i1}, x_{i2}, \dots, x_{ij})$. We use the shorthand notation $v_i = (\tilde{x}_i, y_i, \tilde{z}_i)$ for all observed variables, and v without subscript for the corresponding vectors for n observations. Table 1 in Supplementary Appendix A provides a succinct summary of the notation.

Marginal structural models (Robins et al., 2000; Hernán et al., 2001) are formulated as marginal distributions of potential outcome/counterfactual random variables which are functionally dependent on hypothetical treatment interventions. Letting a_j index r discrete treatment alternatives at time-point j, the r^m potential outcomes for individual i are denoted as $\mathbf{y}_{\tilde{a}i}$, $\tilde{a} \equiv (a_1, \dots, a_m)$. Assuming that the intervention is well-defined and there is no interference between subjects (the *consistency* assumption), the observed outcome is given by $y_i = \sum_{\tilde{a}} \mathbf{1}_{\{\tilde{z}_i = \tilde{a}\}} \mathbf{y}_{\tilde{a}i}$. A marginal structural model then specifies the r^m marginal distributions $p(\mathbf{y}_{\tilde{a}i} \mid \theta)$ through the parameters θ.

Under a data generating mechanism without confounding, the marginal structural model can be estimated using its observed counterpart $p(y_i \mid \tilde{z}_i, \theta)$. Assuming that the *no unmeasured confounding/sequential randomization* condition $\mathbf{y}_{\tilde{a}i} \perp\!\!\!\perp z_{ij} \mid (\tilde{z}_{i(j-1)}, \tilde{x}_{ij})$ and the *positivity* condition $p(z_{ij} = a_j \mid \tilde{z}_{i(j-1)}, \tilde{x}_{ij}) > 0$ hold true for all i, j, and $\tilde{a}$, the parameter θ may be estimated by maximizing the IPT-weighted pseudo-likelihood function

$$q(\theta; v, \gamma, \alpha) \equiv \prod_{i=1}^{n} p(y_i \mid \tilde{z}_i, \theta)^{w_i}, \tag{1}$$

where

$$w_i = \frac{\prod_{j=1}^{m} p(z_{ij} \mid \tilde{z}_{i(j-1)}, \alpha_j)}{\prod_{j=1}^{m} p(z_{ij} \mid \tilde{z}_{i(j-1)}, \tilde{x}_{ij}, \gamma_j)}$$

defines "stabilized" case weights. Here $\alpha \equiv (\alpha_1, \dots, \alpha_m)$ and $\gamma \equiv (\gamma_1, \dots, \gamma_m)$ parametrize the marginal and conditional treatment assignment probabilities, respectively, with the true values of the parameters (γ, α) (for now) taken to be known. The weights w_i in (1) have the property that $E[w_i] = 1$ (see, e.g., Hernán and Robins, 2006, p. 584). This fact does not make (1) a proper likelihood in the sense that the corresponding score variance would equal the Fisher information.

Since the effect of the weighting is to construct a pseudo-population in which there are no imbalances on measured covariates between the treatment groups (Robins et al., 2000, p. 553), (1) can be understood in terms of the relevance weighted likelihood discussed by Hu and Zidek (2002) which arises when a sample from the population of interest is not directly available, but samples from other populations are relevant for learning about this *target* population. Now the target population is one where $z_{ij} \perp\!\!\!\perp \tilde{x}_{ij} \mid \tilde{z}_{i(j-1)}$ holds true; the weights convey information on how much the *observed* population resembles the target population. This information in turn is contained in the parameters γ. In addition, the target population has the same marginal treatment assignment distribution as the observed population, characterized by the parameters α. In the following section we formalize the notion of the target population and relate it to the observed population.

If the true values of the parameters (γ, α) are known, the weights w_i are fixed; to represent random sampling of the original n subjects of equal information contribution, we may consider the likelihood-analogue

$$q(\theta; v, \gamma, \alpha, \pi) = \prod_{i=1}^{n} p(y_i \mid \tilde{z}_i, \theta)^{n\pi_i w_i}, \tag{2}$$

where $\pi \equiv (\pi_1, \dots, \pi_n) \sim \text{Dirichlet}(1, \dots, 1)$, as in the weighted likelihood bootstrap of Newton and Raftery (1994, p.4). An alternative formulation could be obtained by replacing in (2) $n\pi_i$ with $\xi \equiv (\xi_1, \dots, \xi_n) \sim \text{Multinomial}(n; n^{-1}, \dots, n^{-1})$. In Sections 3.2–3.4 we show that randomly drawing vectors $\pi_{(k)}$ (or $\xi_{(k)}$), $k = 1, \dots, l$, and taking $\widehat{\theta}_{(k)} \equiv \arg\max_\theta q(\theta; v, \gamma, \alpha, \pi_{(k)})$ produces an approximate sample of size l from the posterior distribution of θ. In practice, parameters (γ, α) would have to be estimated as well, which we also address below.

3.2. *Bayesian Model Parametrization*

In addition to the variables introduced previously, longitudinal settings often involve latent individual level "frailty" variables, which are determinants of both the outcome and the intermediate variables, but can sometimes be assumed conditionally independent of the treatment assignments. We denote these variables by u_i, and now consider a formal Bayesian construction. We assume that the quadruples $(\tilde{x}_i, y_i, \tilde{z}_i, u_i)$ are infinitely *exchangeable* over the unit indices $i = 1, \dots, n, n + 1, \dots$, and deduce the de Finetti representation (e.g., Bernardo and Smith, 1994, Chapter 4) for the joint distribution of a random sample of size n from such a super-population as

$$\begin{aligned} p(v \mid \mathcal{O}) &= \int_{\phi,\gamma,u} p(\tilde{x}, y, \tilde{z}, u \mid \phi, \gamma, \mathcal{O}) p(\phi, \gamma) \, \mathrm{d}\phi \, \mathrm{d}\gamma \\ &= \int_{\phi,\gamma} \prod_{i=1}^{n} \Bigg[\int_{u_i} p(y_i \mid \tilde{x}_i, \tilde{z}_i, u_i, \phi_1) \\ &\quad \times \prod_{j=1}^{m} p(x_{ij} \mid \tilde{z}_{i(j-1)}, \tilde{x}_{i(j-1)}, u_i, \phi_{2j}) p(u_i \mid \phi_3) \, \mathrm{d}u_i \\ &\quad \times \prod_{j=1}^{m} p(z_{ij} \mid \tilde{z}_{i(j-1)}, \tilde{x}_{ij}, \gamma_j, \mathcal{O}) \Bigg] p(\phi, \gamma) \, \mathrm{d}\phi \, \mathrm{d}\gamma, \end{aligned} \tag{3}$$

assuming that the prior distribution for parameters (ϕ, γ) implied by the representation theorem—presumed here to be finite dimensional for convenience—is absolutely continuous with respect to Lebesgue measure, with density $p(\phi, \gamma)$. Further, $\phi = (\phi_1, \phi_2, \phi_3)$ is a partitioning of ϕ corresponding to the above factorization of the likelihood function, that is, ϕ_1 specifying the conditional outcome model, ϕ_2 the covariate process, and ϕ_3 the marginal distribution of the frailties. The notation $\mathcal{O}$ indexes the data generating mechanism under the observational setting where the treatment assignment can depend on the $\tilde{x}_{ij}$ covariates (cf. Dawid and Didelez, 2010; Røysland, 2011).

Equation (3) follows under the assumption that $z_{ij} \perp\!\!\!\perp u_i \mid (\tilde{x}_{ij}, \tilde{z}_{i(j-1)}, \mathcal{O})$, which is the counterpart of the no unmeasured confounding condition stated in the previous section (cf. Arjas, 2012, Definition 2). The parameter vectors ϕ and γ, specified by the representation theorem as some functions of the infinite sequence of observables, are assumed a priori independent. We note that here ϕ is not of direct interest: what is central to what follows is the interpretation of the parameter vector γ. We define a correctly specified treatment assignment model as the sequence of conditional distributions implied by (3), parameterized via γ. It follows that the outcomes are non-informative about the treatment assignment mechanism, characterized by the parameters γ. To see this, the marginal posterior density for γ may be written

$$\begin{aligned} p(\gamma \mid v, \mathcal{O}) = & \int_{\phi,u} p(\gamma, \phi, u \mid v, \mathcal{O}) \, \mathrm{d}\phi \, \mathrm{d}u \\ & \propto \int_{\phi,u} p(\tilde{x}, y, \tilde{z}, u \mid \phi, \gamma, \mathcal{O}) p(\phi) p(\gamma) \, \mathrm{d}\phi \, \mathrm{d}u \\ & \propto \prod_{i=1}^{n} \prod_{j=1}^{m} p(z_{ij} \mid \tilde{z}_{i(j-1)}, \tilde{x}_{ij}, \gamma_j, \mathcal{O}) p(\gamma) \\ & \propto p(\gamma \mid \tilde{x}, \tilde{z}, \mathcal{O}). \end{aligned} \tag{4}$$

Under the usual regularity assumptions, the posterior in (4) converges to a degenerate distribution at the true value of γ when $n \to \infty$ (cf. van der Vaart, 1998, p. 139).

For causal considerations, we need to envision sampling taking place from another, entirely conceptual, super-population where treatments are assigned *completely at random* so that $z_{ij} \perp\!\!\!\perp (\tilde{x}_{ij}, u_i) \mid (\tilde{z}_{i(j-1)}, \mathcal{E})$, $j = 1, \ldots, m$. The indexing of the probability distributions by $\mathcal{E}$ refers to the characteristics of a conceptual "randomized" version of the treatment assignment mechanism, corresponding to the randomized trial measure considered by Røysland (2011). Causal inferences are then possible if the treatment effect under $\mathcal{E}$ can be estimated based on the data observed under $\mathcal{O}$. In addition, the marginal treatment assignment probabilities under $\mathcal{E}$ are taken to be the same as under the observational setting. The resulting de Finetti representation is

$$\begin{aligned} p(v \mid \mathcal{E}) = \int_{\phi,\alpha} \prod_{i=1}^{n} \Bigg[\int_{u_i} & p(y_i \mid \tilde{x}_i, \tilde{z}_i, u_i, \phi_1) \\ & \times \prod_{j=1}^{m} p(x_{ij} \mid \tilde{z}_{i(j-1)}, \tilde{x}_{i(j-1)}, u_i, \phi_{2j}) p(u_i \mid \phi_3) \, \mathrm{d}u_i \\ & \times \prod_{j=1}^{m} p(z_{ij} \mid \tilde{z}_{i(j-1)}, \alpha_j, \mathcal{E}) \Bigg] p(\phi) p(\alpha) \, \mathrm{d}\phi \, \mathrm{d}\alpha. \qquad (5) \end{aligned}$$

Under standard conditions, the corresponding posterior $p(\alpha \mid \tilde{z}, \mathcal{E})$ converges to a degenerate distribution at the true value of α. An alternative parametrization would be obtained by assuming the pairs $(y_i, \tilde{z}_i)$ to be infinitely exchangeable over the unit indices i. Under the treatment assignment mechanism $\mathcal{E}$ this is sensible, since now the covariates $\tilde{x}_i$ are not confounders and are thus irrelevant to learning about the relationship between the treatment and the outcome. The resulting parametrization is

$$\begin{aligned} p(y, \tilde{z} \mid \mathcal{E}) = \int_{\theta,\alpha} \prod_{i=1}^{n} \Bigg[p(y_i \mid \tilde{z}_i, \theta) \prod_{j=1}^{m} p(z_{ij} \mid \tilde{z}_{i(j-1)}, \alpha_j, \mathcal{E}) \Bigg] \\ \times p(\theta) p(\alpha) \, \mathrm{d}\theta \, \mathrm{d}\alpha. \qquad (6) \end{aligned}$$

The parameters α are the same as in (5), and θ parameterizes the marginal treatment effect of interest. In the Appendix, we motivate the above definitions by linking the representations (3) and (5) to the causal parameter. In order to make causal inferences about θ in (6), one needs to hypothesize generating predictions $v_i^* \equiv (\tilde{x}_i^*, y_i^*, \tilde{z}_i^*)$ from the super-population/data generating mechanism characterized by (5), based on the actually observed sample v of size n from (3). This is in principle straightforward, since

$$p(v_i^* \mid v, \mathcal{E}) = \int_{\phi,\alpha,u_i^*} p(v_i^*, u_i^* \mid \phi, \alpha) p(\phi, \alpha \mid v, \mathcal{E}) \, \mathrm{d}u_i^* \, \mathrm{d}\phi \, \mathrm{d}\alpha$$

where $p(\phi, \alpha \mid v, \mathcal{E}) = p(\phi \mid v) p(\alpha \mid \tilde{z}, \mathcal{E})$, and further

$$\begin{aligned} p(\phi \mid v) \propto \prod_{i=1}^{n} \Bigg[\int_{u_i} & p(y_i \mid \tilde{x}_i, \tilde{z}_i, u_i, \phi_1) \\ & \times \prod_{j=1}^{m} p(x_{ij} \mid \tilde{z}_{i(j-1)}, \tilde{x}_{i(j-1)}, u_i, \phi_{2j}) p(u_i \mid \phi_3) \, \mathrm{d}u_i \Bigg] \\ & \times p(\phi). \end{aligned}$$

However, we wish to avoid specifying the model components parameterized in terms of ϕ, as they reference the latent and unobserved u_i. If, on the other hand, the latent variables are ignored, the modeling approach would be susceptible to the "null paradox" discussed by Robins and Wasserman (1997). We note that our formulation of causal inference as a posterior predictive problem closely resembles the original Bayesian approach by Rubin (1978).

3.3. *IPT Weighting Derived Through a Bayes Decision Rule*

The representations (3) and (5) are linked through the importance sampling identity (e.g., Robert and Casella, 2004, p. 92). Let $U(\cdot)$ be a utility function relevant to the estimation/decision problem. Then

$$\begin{aligned} E[U(v_i^*) \mid v, \mathcal{E}] &= \int_{v_i^*} U(v_i^*) p(v_i^* \mid v, \mathcal{E}) \, \mathrm{d}v^* \\ &= \int_{v_i^*} U(v_i^*) \frac{p(v_i^* \mid v, \mathcal{E})}{p(v_i^* \mid v, \mathcal{O})} p(v_i^* \mid v, \mathcal{O}) \, \mathrm{d}v^* \\ &\equiv \int_{v_i^*} w_i^* \, U(v_i^*) p_n(v_i^*) \, \mathrm{d}v_i^*, \qquad (7) \end{aligned}$$

where p_n is taken to be a non-parametric posterior predictive density in the sense of Walker (2010, p. 26), and $w_i^* = p(v_i^* \mid v, \mathcal{E}) / p(v_i^* \mid v, \mathcal{O})$, which simplifies into

$$\begin{aligned} w_i^* &= \frac{\int_{\alpha} \prod_{j=1}^{m} p(z_{ij}^* \mid \tilde{z}_{i(j-1)}^*, \alpha_j, \mathcal{O}) p(\alpha \mid \tilde{z}, \mathcal{O}) \, \mathrm{d}\alpha}{\int_{\gamma} \prod_{j=1}^{m} p(z_{ij}^* \mid \tilde{z}_{i(j-1)}^*, \tilde{x}_{ij}^*, \gamma_j, \mathcal{O}) p(\gamma \mid \tilde{x}, \tilde{z}, \mathcal{O}) \, \mathrm{d}\gamma} \\ &= \frac{E_{\alpha} \left[\prod_{j=1}^{m} p(z_{ij}^* \mid \tilde{z}_{i(j-1)}^*, \alpha_j, \mathcal{O}) \mid \tilde{z}, \mathcal{O} \right]}{E_{\gamma} \left[\prod_{j=1}^{m} p(z_{ij}^* \mid \tilde{z}_{i(j-1)}^*, \tilde{x}_{ij}^*, \gamma_j, \mathcal{O}) \mid \tilde{x}, \tilde{z}, \mathcal{O} \right]}, \qquad (8) \end{aligned}$$

an estimated version of the weight in (1). The form (7) is expressed entirely in terms of observable quantities, since $p(z_{ij} \mid \tilde{z}_{i(j-1)}, \alpha_j, \mathcal{E}) = p(z_{ij} \mid \tilde{z}_{i(j-1)}, \alpha_j, \mathcal{O})$. In (7), we require that the ratio $p(v_i^* \mid v, \mathcal{E}) / p(v_i^* \mid v, \mathcal{O})$ is well-defined (formally, we require absolute continuity of the experimental measure with respect to the observational measure, cf. Dawid and Didelez, 2010, p. 196). This implies in particular that the treatment assignments z_{ij} under $\mathcal{O}$ may not be deterministic, and is the counterpart of the positivity condition (see, e.g., Hernán and Robins, 2006, pp. 582–583). The no unmeasured confounding condition $z_{ij} \perp\!\!\!\perp u_i \mid (\tilde{x}_{ij}, \tilde{z}_{i(j-1)}, \mathcal{O})$ is also required for obtaining the simplified form (8), the terms involving the latent variables u_i canceling out of the fraction.

Based on (6), we choose the utility function $U(v_i^*; \theta) \equiv \log p(y_i^* \mid \tilde{z}_i^*, \theta) \equiv \ell(y_i^* \mid \tilde{z}_i^*, \theta)$ say, and then maximize the expected utility with respect to the parameters of interest θ. Following Walker (2010, p. 27) and adopting the Bayesian bootstrap strategy $p_n(v_i^*) = \sum_{k=1}^n \pi_k \delta_{v_k}(v_i^*)$ where $\pi \equiv (\pi_1, \ldots, \pi_n) \sim \text{Dirichlet}(1, \ldots, 1)$, we then obtain the log-likelihood-analogue corresponding to (2) through

$$\begin{aligned} E\left[\ell(y_i^* \mid \tilde{z}_i^*, \theta) \mid v, \mathcal{E}\right] &= \int_{v_i^*} w_i^* \ell(y_i^* \mid \tilde{z}_i^*, \theta) \sum_{k=1}^n \pi_k \delta_{v_k}(v_i^*)\, \mathrm{d}v_i^* \\ &= \sum_{i=1}^n \pi_i w_i \ell(y_i \mid \tilde{z}_i, \theta). \end{aligned} \tag{9}$$

Consequently,

$$\begin{aligned} &\arg\max_\theta E\left[\ell(y_i^* \mid \tilde{z}_i^*, \theta) \mid v, \mathcal{E}\right] \\ &\quad = \arg\max_\theta \left[\sum_{i=1}^n \pi_i w_i \ell(y_i \mid \tilde{z}_i, \theta)\right] \equiv \widehat{\theta}(v; \pi), \end{aligned} \tag{10}$$

the weighted maximum likelihood estimator of θ.

3.4. *A Computational Algorithm*

As in Newton and Raftery (1994), an approximate sample from the posterior distribution of θ may now be produced by taking a sample $(\pi_{(1)}, \ldots, \pi_{(l)})$ of the weight vectors of length n from the uniform Dirichlet distribution, and taking $(\theta_{(1)}, \ldots, \theta_{(l)}) = (\widehat{\theta}(v; \pi_{(1)}), \ldots, \widehat{\theta}(v; \pi_{(l)}))$ to be a sample from $p(\theta \mid v, \mathcal{E})$. Alternatively, π could be replaced by the multinomial random vector ξ. It should be noted that the weighted log-likelihood function (9) cannot be used in place of a likelihood function in Bayes' formula, and its curvature does not play a direct part in quantifying the uncertainty on θ. This estimation approach as such does not allow specifying an informative (non-flat) prior on θ. However, if required, informative priors could be incorporated using the sampling-importance resampling (SIR, Rubin, 1988) approach as discussed by Newton and Raftery (1994). In short, in this procedure a (say, kernel) density estimate g would be calculated from the initial sample $(\theta_{(1)}, \ldots, \theta_{(l)})$, followed by resampling with the importance weights $L(\theta_{(k)})p(\theta_{(k)})/g(\theta_{(k)})$, where L is a likelihood function and p is the informative prior. In the present setting we do not have a closed form likelihood function, but the posterior density estimate g under flat priors can be taken as a numerical likelihood, resulting in importance resampling weights $p(\theta_{(k)})$. Alternatively, to avoid potential issues in the importance resampling weights, the numerical likelihood g may be used directly in the Bayes' formula in place of a closed form likelihood function, enabling the use of standard Markov chain Monte Carlo (MCMC) methods, and informative prior specifications for θ. We illustrate this augmented procedure in Supplementary Appendix B.

Prior specifications for γ and α and posterior inferences from $p(\gamma \mid \tilde{x}, \tilde{z}, \mathcal{O})$ and $p(\alpha \mid \tilde{z}, \mathcal{O})$ proceed in the usual way, the evaluation of the weights (8) using Monte Carlo integration requiring only a single MCMC sample from these posteriors. We note that when there is no confounding under the observational setting, that is, $z_{ij} \perp\!\!\!\perp \tilde{x}_{ij} \mid (\tilde{z}_{i(j-1)}, \mathcal{O})$, $j = 1, \ldots, m$, the weights $w_i \to 1$ and the estimator coincides asymptotically with the unweighted maximum likelihood estimator.

The proposed computational algorithm can be summarized as follows: first the treatment assignment model is fitted using standard Bayesian MCMC techniques to obtain the posterior mean treatment assignment probabilities and IPT weights. Second, an approximate sample is produced from the posterior distribution of the MSM parameters θ with flat priors by fitting the MSM using a Bayesian bootstrap procedure where the obtained IPT weights are multiplied by uniform Dirichlet resampling weights. The procedure can be augmented to accommodate informative priors for θ. A step-by-step representation of the computational algorithm is given in Supplementary Appendix B.

4. Previously Proposed Two-Step and Joint Bayesian Estimation Approaches

4.1. *Two-Step Estimation*

Previous Bayesian approaches proposed by Hoshino (2008) and Kaplan and Chen (2012) for Bayesian propensity score adjustment or weighting are implicitly based on a marginal quasi-posterior distribution of the form

$$q(\theta; v) \equiv \int_\gamma q(\theta; v, \gamma) p(\gamma \mid \tilde{x}, \tilde{z})\, \mathrm{d}\gamma. \tag{11}$$

The quasi-Bayes point estimator of Hoshino (2008) would be obtained as the mean of (11), in practice evaluated using MCMC sampling where the likelihood is replaced by the IPT-weighted pseudo-likelihood. Given a sample $\gamma_{(k)}$, $k = 1, \ldots, l$ from $p(\gamma \mid \tilde{x}, \tilde{z})$, the multiple imputation type point estimator of Kaplan and Chen (2012), also implied by (11), is $E_{\gamma \mid \tilde{x}, \tilde{z}}[E(\theta \mid v, \gamma)] \approx \frac{1}{l} \sum_{k=1}^l \widehat{\theta}(v; \gamma_{(k)})$. Such point estimators are consistent as, under standard regularity conditions, $p(\gamma \mid \tilde{x}, \tilde{z})$ converges to a point mass at the truth. However, since $q(\theta; v; \gamma)$ is not a likelihood, the integral $q(\theta; v)$ does not have a probabilistic interpretation. In particular, since (11) is not a true posterior distribution, it does not readily provide a mechanism for variance estimation. We refer to Supplementary Appendix C for more details.

4.2. *Joint Estimation*

Approaches to Bayesian (and likelihood-based) propensity score adjustment which allow feedback between the outcome model and the treatment assignment model have been a source of continuing controversy in the literature (e.g., McCandless et al., 2010; Kaplan and Chen, 2012; Zigler et al., 2013). Results from Section 3.2 give insight into this issue; we elaborate in Supplementary Appendix C. Briefly, we conclude that many of the proposed joint estimation methods are not true propensity score adjustment methods in the sense that they do not retain the balancing property of propensity scores.

Table 1

Results for point and variance estimators of θ_2 over 1000 replications. The columns correspond to estimator, mean point estimate, bias relative to the true value of θ_2 (RB), Monte Carlo standard deviation of the point estimates (SD), mean standard error estimate (SE), standard error estimate bias relative to the Monte Carlo SD, and 95% confidence interval coverage probability (CP).

Scenario	Estimator	Mean	RB (%)	SD	SE	RB (%)	95% CP
$b = 0$,	Naive	−0.252	−1.991	0.106	0.106	−0.413	95.1
$\theta_2 = -0.247$	ITPW, sandwich	−0.253	−2.179	0.107	0.107	−0.121	95.8
	ITPW, Adj. sandwich	−0.253	−2.179	0.107	0.105	−2.058	95.2
	quasi-Bayes	−0.255	−3.018	0.109	0.104	−4.559	93.9
	MI	−0.253	−2.421	0.108	0.113	4.676	96.6
	Bayes/Dirichlet	−0.257	−3.752	0.108	0.108	−0.717	95.5
	Bayes/Multinomial	−0.257	−3.806	0.108	0.109	0.370	94.8
	Bootstrap	−0.257	−3.801	0.108	0.109	0.578	95.0
$b = 0.15$,	Naive	−0.345	39.456	0.122	0.124	1.823	52.6
$\theta_2 = -0.569$	ITPW, sandwich	−0.570	−0.141	0.142	0.142	−0.272	95.0
	ITPW, Adj. sandwich	−0.570	−0.141	0.142	0.133	−6.434	93.7
	quasi-Bayes	−0.587	−3.080	0.147	0.147	0.379	95.4
	MI	−0.582	−2.266	0.145	0.159	9.449	97.7
	Bayes/Dirichlet	−0.576	−1.102	0.143	0.141	−0.937	94.6
	Bayes/Multinomial	−0.576	−1.134	0.142	0.144	1.088	95.4
	Bootstrap	−0.577	−1.430	0.143	0.141	−0.901	95.0
$b = 0.3$,	Naive	−0.184	76.340	0.124	0.127	2.540	0.8
$\theta_2 = -0.777$	ITPW, sandwich	−0.757	2.591	0.217	0.198	−8.750	93.5
	ITPW, Adj. sandwich	−0.757	2.591	0.217	0.174	−19.665	90.0
	quasi-Bayes	−0.795	−2.325	0.230	0.284	23.717	97.0
	MI	−0.789	−1.540	0.229	0.236	3.258	97.7
	Bayes/Dirichlet	−0.755	2.888	0.207	0.191	−7.750	93.3
	Bayes/Multinomial	−0.754	3.021	0.204	0.200	−1.892	94.8
	Bootstrap	−0.759	2.398	0.206	0.195	−5.322	93.2

5. Marginal Structural Model: Simulation Study

5.1. *Simulation Strategy*

Algorithms for simulating outcomes from a given marginal structural model are available (e.g., Havercroft and Didelez, 2012) and can be used to deduce the marginal parameters of interest even in the presence of mediation and noncollapsibility by appealing to standard Monte Carlo principles. Here, following Section 3.2, we do not regard marginal structural models as data generating mechanisms as such, but instead define θ to be a parameter of a given regression model $p(y_i \mid \tilde{z}_i, \theta)$ fitted to an infinite sequence of observations from a data generating mechanism characterized by the representation (5) (cf. Gelman, 2007, pp. 157–158). In the Appendix we show that (5) is fully specified by (3). The limiting value of θ as $n \to \infty$ is thus fully defined by the distributions in (3) and a given model specification $p(y_i \mid \tilde{z}_i, \theta)$, and is here taken to be the quantity of interest. The correct marginal model is specified by (12) in Appendix, but under mild regularity conditions the limiting value if θ exists irrespective of whether the postulated model is correct (cf. White, 1982), and can be approximated up to arbitrary precision by simulation.

We approximate the limiting value of θ by simulating the r^m potential outcomes for each $i = 1, \ldots, N$, $N \gg n$, from (3) and fitting the marginal model to the resulting Nr^m observations. In the data generating mechanism we choose $m = 3$ time intervals, $r = 2$ treatment levels, 5 covariates and $n = 500$. The conditional distributions in (3) for our three interval MSM simulation study are given in the Appendix. We considered three different scenarios with increasing degree of confounding, corresponding to $b = 0$, $b = 0.15$, and $b = 0.3$.

5.2. *Simulation Study: Results*

The fitted treatment assignment models were chosen as $\text{logit}\{p(z_{ij} \mid \tilde{z}_{i(j-1)}, \tilde{x}_{ij}, \gamma_j)\} = \gamma_{j1} + \gamma_{j2}^{\top}\tilde{z}_{i(j-1)} + \gamma_{j3}^{\top}\tilde{x}_{ij}$ and $\text{logit}\{p(z_{ij} \mid \tilde{z}_{i(j-1)}, \alpha_j)\} = \alpha_{j1} + \alpha_{j2}^{\top}\tilde{z}_{i(j-1)}$ for $j = 1, 2, 3$, and the marginal model as $\text{logit}\{p(y_i \mid \tilde{z}_i, \theta)\} = \theta_1 + \theta_2 \sum_{j=1}^{3} z_{ij}$. The results over 1000 simulation rounds for several point estimators are presented in Table 1. In particular, we consider (i) the naive unweighted estimator which does not account for confounding; (ii) the typical, frequentist IPT weighted estimator ("IPTW"), with the plug-in estimates $(\widehat{\gamma}, \widehat{\alpha})$ substituted in (1); (iii) the quasi-Bayes estimator (Hoshino, 2008) given by the mean of the marginal quasi-posterior distribution (11); (iv) the corresponding multiple imputation type point estimator ("MI"); (v) our proposed Bayesian approach with Dirichlet sampling ("Bayes/Dirichlet"); (vi) our proposed Bayesian approach with Multinomial sampling ("Bayes/Multinomial"); and (vii) the estimate based on bootstrapping the frequentist IPT weighted estimator, where the treatment assignment models are re-fitted and weights re-calculated in each bootstrap sample. The weights in estimators (v) and (vi) were based on MCMC samples from the posterior distributions of γ and α, using flat improper priors for these parameters. Estimators (v)–(vii) were calculated from 2500 replications.

The results show that all of the weighted estimators are approximately unbiased, although in the second and third scenarios the estimators (iii) and (iv) based on (11) give slightly different results from the other estimators, both in terms of bias and excess variability. In the last scenario, the resampling-based point estimators (v)–(vii) show slightly lower variability than the standard IPTW estimator (i); this is due to the most influential observations not being present in every resample. This suggests that when large weights are present, resampling might be useful for improving the stability of point estimation.

The standard approach for variance estimation of IPTW-based estimators is the "robust"/sandwich variance estimator, which is expected to be conservative when the nuisance parameters are fixed to their maximum likelihood estimates (Hernán et al., 2001, p. 444). Since the asymptotic variance of the IPT-weighted estimator with estimated weights (at $(\widehat{\gamma}, \widehat{\alpha})$) is smaller than that of the same estimator with the true weights (at the true values of (γ, α); see, e.g., Henmi and Eguchi, 2004), a Taylor expansion-based correction term may be subtracted from the sandwich estimator to account for the estimation of the weights (e.g., Robins, Mark, and Newey, 1992). However, it is also well known that the sandwich estimator itself is often biased downwards in small samples (e.g., Fay and Graubard, 2001). This is more pronounced when influential observations with large weights are present, and thus correcting the sandwich estimator downwards in such situations may not be sensible.

Table 1 gives also the estimated standard errors for each of the point estimators. The 95% confidence interval coverage probabilities correspond to normal approximation confidence intervals calculated using the respective variance estimates, except for the Multinomial/Dirichlet sampling and bootstrap estimators, for which we report the sampling/posterior distribution based confidence intervals. The results under the second and third scenarios indicate that adjustment for estimation of γ and α may indeed adversely affect the small sample properties of the sandwich variance estimator, which itself shows underestimation when $b = 0.3$. The quasi-posterior variances are not appropriate for variance estimation, and this seems to be the case also for the multiple imputation type variance decomposition. The Bayesian estimators do reasonably well under all three scenarios, giving results similar to the frequentist bootstrap. We also repeated the simulations with $n = 1000$ and $n = 2000$ (see Supplementary Appendix D), with the conclusions essentially unchanged.

The simulations demonstrate that the variance estimators which rely on asymptotic approximations—the sandwich estimator and its adjusted version—have a tendency for underestimation under settings where influential observations with large weights are present. The proposed Bayesian approach with Dirichlet sampling seems to be less affected by the presence of influential observations.

6. ART Interruption and Liver Fibrosis in HIV/HCV Co-Infected Individuals

6.1. *Study Background*

We now revisit the real data example introduced in Section 2. We update an earlier analysis of Thorpe et al. (2011), as the cohort has since been followed up for nearly two additional years, increasing the number of outcome events from 53 to 112. Similar criteria as in Thorpe et al. (2011) were used to select individuals into the analysis; we included co-infected adults who were not on HCV treatment and did not have liver fibrosis at baseline, according to the outcome definition below. Individuals suspected of having spontaneously cleared their HCV infection (based on two consecutive negative HCV viral load measurements) were excluded as they are not considered at risk for fibrosis progression. The outcome event was defined as aminotransferase-to-platelet ratio index (APRI) being at least 1.5 in any subsequent visit, this event being a surrogate marker for liver fibrosis. We included visits where the individuals were either on ART ($z_{ij} = 0$) or had interrupted therapy ($z_{ij} = 1$), during the 6 months before each follow-up visit. To ensure correct temporal order in the analyses, in the treatment assignment model all time-varying covariates (x_{ij}), including the laboratory measurements (HIV viral load and CD4 cell count), were lagged one visit. Follow-up was terminated at the outcome event ($y_{ij} = 1$); individuals starting HCV medication during the follow-up were censored. These selections resulted in $N = 474$ individuals with at least one follow-up visit (scheduled at every 6 months) after the baseline visit, and 2066 follow-up visits in total (1592 excluding the baseline visits). The number of follow-up visits m_i ranged from 2 to 16 (median 4).

6.2. *Analysis*

Our main objectives are to compare the variance estimates given by the alternative methods under a real setting, as well as to demonstrate that the approach in Section 3.2 generalizes to longitudinal settings with censoring. The details on accommodating censoring to the weighting approach of Section 3 are given in Supplementary Appendix E. In short, in addition to the marginal and conditional treatment assignment models, specified as pooled logistic regressions $\text{logit}\{P(z_{ij} = 1 \mid z_{i(j-1)}, \alpha)\} = \alpha z_{i(j-1)}$ and $\text{logit}\{P(z_{ij} = 1 \mid z_{i(j-1)}, x_{i(j-1)}, \gamma)\} = \gamma^{\top}(z_{i(j-1)}, x_{i(j-1)})$, $j = 2, \ldots, m_i$, we need to estimate marginal and conditional censoring models $\text{logit}\{P(c_{ij} = 1 \mid z_{ij}, \mu)\} = \mu z_{ij}$ and $\text{logit}\{P(c_{ij} = 1 \mid z_{ij}, x_{ij}, \eta)\} = \eta^{\top}(z_{ij}, x_{ij})$, $j = 1, \ldots, m_i - y_{im_i}$. The potential confounders we considered were baseline covariates female gender, hepatitis B surface antigen (HBsAg) test and baseline APRI, as well as time-varying covariates age, current intravenous drug use (binary), current alcohol use (binary), duration of HCV infection, HIV viral load, CD4 cell count, as well as ART interruption status at the previous visit. The conditional model estimates are shown in Table 2. The maximum stabilized visit specific cumulative weight calculated at the MLEs $(\widehat{\eta}, \widehat{\mu}, \widehat{\gamma}, \widehat{\alpha})$ was only 2.95; this is due to lagged interruption being the only significant predictor of present interruption (Table 2). With little variability in the weights, the results for the alternative estimators would be expected to follow the pattern in the first simulation scenario.

Due to the binary outcome status determined at each follow-up visit (as opposed to once at the end of the follow-up) and the relatively low rate of events, we used pooled logistic regression $\text{logit}\{p(y_{ij} = 1 \mid z_{ij}, \theta)\} = \theta_1 + \theta_2 z_{ij}$ as the specification for the MSM. Table 3 shows the estimates for the interruption effect θ_2 in the marginal model and the corresponding

Table 2
Maximum likelihood estimates from pooled logistic regression for the ART interruption exposure and censoring at end of the follow-up in the CCC data

	Current interruption			Censoring		
Covariate	MLE	SE	z	MLE	SE	z
Lagged interruption	4.616	0.333	13.853	0.039	0.256	0.151
Female gender	0.557	0.304	1.833	0.163	0.134	1.222
Log baseline APRI	0.060	0.290	0.208	−0.097	0.114	−0.852
HBsAg	0.382	0.879	0.434	0.352	0.326	1.080
Age	−0.012	0.019	−0.626	0.018	0.008	2.347
CD4 cell count/100	0.001	0.052	0.029	0.035	0.018	1.909
Log HIV RNA	0.084	0.055	1.522	−0.009	0.032	−0.287
Intravenous drug use	−0.148	0.310	−0.477	−0.061	0.132	−0.464
Current alcohol use	0.108	0.291	0.372	−0.078	0.119	−0.660
HCV duration	0.010	0.016	0.635	0.006	0.006	0.960

standard errors. The weights in the Bayesian estimators were calculated from MCMC samples from the posterior distributions of $(\eta, \mu, \gamma, \alpha)$ using flat improper priors. Multinomial, Dirichlet and bootstrap estimates were calculated from 2500 replications. The five alternative estimates are similar, with the exception of the MI-type estimator, which, as in the simulations, appears to overestimate the standard error. In contrast, the Multinomial and Dirichlet sampling standard errors are close to the bootstrap one, without involving re-estimation of the treatment and censoring models in each replication.

7. Discussion

In attempts to incorporate variability due to estimation of the propensity scores or IPT weights into Bayesian inferences of treatment effects, it has not always been recognized that from the frequentist point of view, estimation of the nuisance models does not add variability to the treatment effect estimate. In addition, standard Bayesian arguments based on exchangeability and de Finetti representations cannot justify outcome model specifications which are functions of the treatment assignment probabilities, unless it is explicitly acknowledged that the model thus specified is also misspecified. In this article, we motivated IPT weighting through a Bayesian decision-theoretic argument, formalizing the notion of pseudo-population which has often been given as an intuitive explanation of the function of IPT weighting (e.g., Joffe et al., 2004).

Table 3
Estimates for the marginal effect of ART interruption (log-hazard ratio) θ_2 on liver fibrosis outcome in the CCC data. Resampling-based estimates are calculated from 2500 replications.

Estimator	$\widehat{\theta}_2$	SE	z
Naive	0.452	0.354	1.278
IPTW, sandwich	0.354	0.377	0.937
MI	0.316	0.529	0.597
Bayes/Dirichlet	0.366	0.375	0.976
Bayes/Multinomial	0.361	0.400	0.902
Bootstrap	0.308	0.395	0.780

We proposed a fully Bayesian approach to estimating parameters of a marginal structural model, formulating the causal inference problem as a Bayesian prediction problem. Our development suggests that the IPT weights should be fixed to values given by the posterior predictive treatment assignment probabilities. The estimated weights then function as importance sampling weights in predicting the outcome in a hypothetical population without covariate imbalances. Our exposition should make significant steps toward resolving the lingering question of whether and how the uncertainty in estimation of weights should be incorporated in Bayesian estimation of marginal treatment effects. Furthermore, our development should motivate further research into the use of non-parametric Bayesian regression and model selection/averaging techniques in estimation of the IPT weights.

8. Supplementary Materials

Supplementary Web Appendices, referenced in Sections 3, 4, 5, and 6, as well as the code for producing the simulation results, are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The majority of this work was carried out when the first author was working at the Department of Epidemiology, Biostatistics and Occupational Health of McGill University. The research of Dr. Saarela was supported by the Finnish Foundation for Technology Promotion. Drs. Saarela, Stephens and Moodie acknowledge support of the Natural Sciences and Engineering Research Council (NSERC) of Canada. Drs. Klein and Moodie are supported by, respectively, a Chercheur-National career award and a Chercheur-Boursier junior 2 career award from the Fonds de recherche du Québec-Santé (FRQ-S). The Canadian HIV/HCV Co-infection Cohort was funded by Réseau SIDA/maladies infectieuses of the FRQ-S, the Canadian Institutes of Health Research (CIHR, MOP-79529) and the CIHR Canadian HIV Trials Network (CTN222).

References

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

Arjas, E. (2012). Causal inference from observational data: A Bayesian predictive approach. In *Causality: Statistical Perspectives and Applications*, C. Berzuini, A. P. Dawid, and L. Bernardinelli (eds), 71–84. New York: Wiley.

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

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

Fay, M. P. and Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* **57**, 1198–1206.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science* **31**, 4190–4206.

Havercroft, W. G. and Didelez, V. (2012). Simulating from marginal structural models with time-dependent confounding. *Statistics in Medicine* **31**, 4190–4206.

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.

Hu, F. and Zidek, J. V. (2002). The weighted likelihood. *The Canadian Journal of Statistics* **30**, 347–371.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., and Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models: Review and new applications. *The American Statistician* **58**, 272–279.

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

Klein, M. B., Saeed, S., Yang, H., Cohen, J., Conway, B., Cooper, C., Côte, P., Cox, J., Gill, J., Haase, D., Haider, S., Montaner, J., Pick, N., Rachlis, A., Rouleau, D., Sandre, R., Tyndall, M., and Walmsley, S. (2010). Cohort profile: The Canadian HIV-Hepatitis C Co-infection Cohort Study. *International Journal of Epidemiology* **39**, 1162–1169.

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**, Article 16.

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

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.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods.* New York: Springer.

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., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48**, 479–495.

Robins, J. M. and Wasserman, L. (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, Providence Rhode Island, August 1–3*, D. Geiger and P. Shenoy (eds), 409–420. San Francisco: Morgan Kaufmann.

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. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* **6**, 34–58.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics 3*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (eds), 395–402. Oxford: Oxford University Press.

Thorpe, J., Saeed, S., Moodie, E. E. M., and Klein, M. B. (2011). Antiretroviral treatment interruption leads to progression of liver fibrosis in HIV-hepatitis C virus co-infection. *AIDS* **25**, 967–664.

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

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

Wang, X. (2006). Approximating Bayesian inference by weighted likelihood. *The Canadian Journal of Statistics* **34**, 279–298.

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

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.

Received October 2013. *Revised August* 2014.
Accepted August 2014.

Appendix

I. Linking the experimental and observational representations. We link the representations (3) and (5) to the causal parameter θ in (6). We note first that (5) is obtained from (3) by noting that $p(z_{ij} \mid \tilde{z}_{i(j-1)}, \alpha_j, \mathcal{E})$ can be written as

$$\frac{\displaystyle\int_{\tilde{x}_{ij}} \prod_{j'=1}^{j} p(z_{ij'} \mid \tilde{z}_{i(j'-1)}, \tilde{x}_{ij'}, \gamma_{j'}, \mathcal{O}) \int_{u_i} I(\tilde{x}_{ij}, u_i)\, \mathrm{d}u_i\, \mathrm{d}\tilde{x}_{ij}}{\displaystyle\int_{\tilde{x}_{ij}} \prod_{j'=1}^{j-1} p(z_{ij'} \mid \tilde{z}_{i(j'-1)}, \tilde{x}_{ij'}, \gamma_{j'}, \mathcal{O}) \int_{u_i} I(\tilde{x}_{ij}, u_i)\, \mathrm{d}u_i\, \mathrm{d}\tilde{x}_{ij}},$$

where

$$I(\tilde{x}_{ij}, u_i) \equiv \prod_{j'=1}^{j} p(x_{ij'} \mid \tilde{z}_{i(j'-1)}, \tilde{x}_{i(j'-1)}, u_i, \phi_{2j'})\, p(u_i \mid \phi_3).$$

Now the outcome model in (6), $p(y_i \mid \tilde{z}_i, \theta)$, is specified by (5) as

$$\frac{\displaystyle\int_{\tilde{x}_i, u_i} p(y_i \mid \tilde{x}_i, \tilde{z}_i, u_i, \phi_1) I(\tilde{x}_i, u_i)\, \mathrm{d}u_i\, \mathrm{d}\tilde{x}_i}{\displaystyle\int_{\tilde{x}_i, u_i} I(\tilde{x}_i, u_i)\, \mathrm{d}u_i\, \mathrm{d}\tilde{x}_i}, \tag{12}$$

where

$$I(\tilde{x}_i, u_i) \equiv \prod_{j=1}^{m} p(x_{ij} \mid \tilde{z}_{i(j-1)}, \tilde{x}_{i(j-1)}, u_i, \phi_{2j}) p(u_i \mid \phi_3).$$

Notably (12) does not depend on α. This is important for the characterization of θ as a causal parameter, as the corresponding marginal distribution under $\mathcal{O}$ would depend on γ.

II. Simulation study. We generate $u_i \sim N_5(0, \Sigma_u)$, and then

1. $x_{i1} \sim N_5(0, \Sigma_x)$; $\quad \text{logit}\{p(z_{i1} = 1 \mid x_{i1})\} = -0.1 + b^\top x_{i1}$
2. $x_{i2} \mid z_{i1}, x_{i1}, u_i \sim N_5\left(x_{i1} - 0.75 z_{i1} + u_i, \frac{1}{16}\Sigma_x\right);$

   $\text{logit}\{p(z_{i2} = 1 \mid z_{i1}, \tilde{x}_{i2})\} = -0.1 + 2z_{i1} + b^\top x_{i2}$
3. $x_{i3} \mid \tilde{z}_{i2}, \tilde{x}_{i2}, u_i \sim N_5\left(x_{i2} - 0.75 z_{i2} + u_i, \frac{1}{16}\Sigma_x\right);$

   $\text{logit}\{p(z_{i3} = 1 \mid \tilde{z}_{i2}, \tilde{x}_{i3})\} = -0.1 + 2z_{i2} + b^\top x_{i3};$

   $\text{logit}\{p(y_i = 1 \mid \tilde{x}_i, \tilde{z}_i, u_i)\}$

   $$= -0.1 - 0.25 \sum_{j=1}^{3} z_{ij} + b^\top x_{i3} + 1^\top u_i/5,$$

where b is a constant vector of length 5, and Σ_x (Σ_u) is a 5×5 covariance matrix with diagonal elements set to 1 (0.1) and off-diagonal elements set to 0.25 (0.05).