Biometrics 71, 279–301
June 2015

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 $i = 1, \ldots, n$, with measurements of covariates and subsequent treatment decisions carried out at discrete time points $j = 1, \ldots, m$. Let $\tilde{z}_i \equiv (z_{i1}, z_{i2}, \ldots, z_{im})$ 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 $\tilde{x}_i \equiv (x_{i1}, x_{i2}, \ldots, x_{im})$ 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, $\tilde{x}_{ij} \equiv (x_{i1}, x_{i2}, \ldots, x_{ij})$. 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 $\mathbf{y}_{\tilde{a}i}$, $\tilde{a} \equiv (a_1, \ldots, a_m)$. 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 $\alpha \equiv (\alpha_1, \ldots, \alpha_m)$ and $\gamma \equiv (\gamma_1, \ldots, \gamma_m)$ 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 $\pi \equiv (\pi_1, \ldots, \pi_n) \sim \text{Dirichlet}(1, \ldots, 1)$, 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 $\xi \equiv (\xi_1, \ldots, \xi_n) \sim \text{Multinomial}(n; n^{-1}, \ldots, n^{-1})$. In Sections 3.2–3.4 we show that randomly drawing vectors $\pi_{(k)}$ (or $\xi_{(k)}$), $k = 1, \ldots, l$, 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 $i = 1, \ldots, n, n+1, \ldots$, 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

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

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

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

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

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

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

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

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

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

$$= \sum_{i=1}^n \pi_i w_i \ell(y_i \mid \tilde{z}_i, \theta). \tag{9}$$

Consequently,

$$\arg\max_\theta E\left[\ell(y_i^* \mid \tilde{z}_i^*, \theta) \mid v, \mathcal{E}\right]$$

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

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 $E_{\gamma|\tilde{x},\tilde{z}}[E(\theta \mid v, \gamma)] \approx \frac{1}{l} \sum_{k=1}^l \widehat{\theta}(v; \gamma_{(k)})$. 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 non-collapsibility 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

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

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. $x_{i1} \sim N_5(0, \Sigma_x)$; $\quad \mathrm{logit}\{p(z_{i1} = 1 \mid x_{i1})\} = -0.1 + b^\top x_{i1}$
2. $x_{i2} \mid z_{i1}, x_{i1}, u_i \sim N_5\left(x_{i1} - 0.75 z_{i1} + u_i, \frac{1}{16}\Sigma_x\right)$;
   $\mathrm{logit}\{p(z_{i2} = 1 \mid z_{i1}, \tilde{x}_{i2})\} = -0.1 + 2 z_{i1} + b^\top x_{i2}$
3. $x_{i3} \mid \tilde{z}_{i2}, \tilde{x}_{i2}, u_i \sim N_5\left(x_{i2} - 0.75 z_{i2} + u_i, \frac{1}{16}\Sigma_x\right)$;
   $\mathrm{logit}\{p(z_{i3} = 1 \mid \tilde{z}_{i2}, \tilde{x}_{i3})\} = -0.1 + 2 z_{i2} + b^\top x_{i3}$;
   $\mathrm{logit}\{p(y_i = 1 \mid \tilde{x}_i, \tilde{z}_i, u_i)\}$
   $= -0.1 - 0.25 \sum_{j=1}^{3} z_{ij} + b^\top x_{i3} + 1^\top u_i / 5,$

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).

Discussions

Michael R. Elliott and Roderick J. Little*
Department of Biostatistics, School of Public Health, University of Michigan,
Ann Arbor, Michigan 48109–2029, U.S.A.
**email:* rlittle@umich.edu

We applaud Saarela, Stephens, Moodie, and Klein (SSMK) for building bridges between Bayesian and frequentist forms of inference for marginal structural models (MSMs), given the tendency of our profession to bifurcate into these sometimes antagonistic camps. SSMK's approach has a strongly Bayesian flavor, and in their simple simulation it shows some promise of generating inferences with superior frequentist properties. However, the formulation is general and complex, and we are not sure to what extent the analysis is fully Bayesian. We consider here some basic examples, which we think help to clarify differences between inverse probability weighting (IPW) and Bayesian approaches to inference, particularly concerning the role of selection and treatment allocation weights. Specifically, SSMK include weights in the Bayesian analysis via the weighted likelihood bootstrap (Newton and Raftery, 1994), an ingenious computational device for generating approximate draws from the posterior distribution; we show that, in our simple examples, weighting the cases is unnecessary for causal inferences from a well-specified Bayesian MSM. However, weights can play an important role as covariates to generate robust model-based inferences.

Example 1: Survey weighting. The first example does not concern MSMs, but it is fundamental, and it sets the stage for our other two examples. A standard weighting approach in sample surveys weights the sampled cases by the inverse of their probabilities of selection. Figure 1A displays data on a population of N units, where X is a set of design variables known for all units in the population, I is an inclusion indicator taking value 1 for sampled units, say $i = 1, \ldots, n$, and 0 for non-sampled units $i = n+1, \ldots, N$. The variables Y are recorded in the sample and hence missing for non-sampled units. The sampling weight is the inverse of the selection probability, which is a function of the design variables, leading to the weight variable $W = W(X)$. A basic design-based (frequentist) approach to inference weights the sampled case i by w_i; for example, the estimate of the mean of Y is $\sum_{i=1}^{n} w_i y_i / N$. From a model-based (and in particular, Bayesian) perspective, a model is defined relating Y to the covariates X, and then non-sampled values of Y are replaced by predictions from the regression of Y on X. In a Markov Chain Monte Carlo simulation of the posterior predictive distribution of a function $g(y_1, \ldots, y_N)$ of Y, such as the population mean, the non-sampled values are simulated using draws $(y_{n+1}^{(d)}, \ldots, y_N^{(d)})$ from their posterior predictive distribution given the sample data $(y_i, x_i), i = 1, \ldots, n$ and $x_i, i = n+1, \ldots, N$. Then $g(y_1, \ldots, y_n, y_{n+1}^{(d)}, \ldots, y_N^{(d)})$ is a draw from the posterior distribution of $g(y_1, \ldots, y_N)$. The draws $y_i^{(d)}$ are predictions from Bayesian regression of Y on X, and

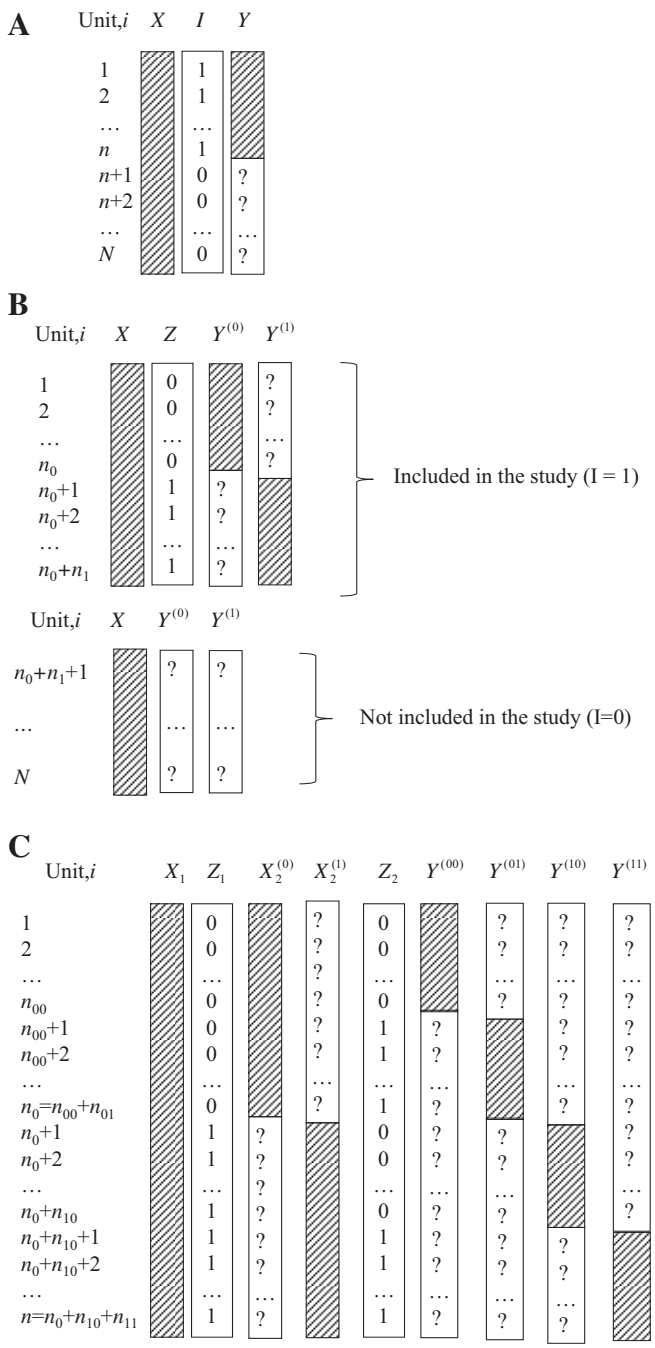

Figure 1. (A) Data Pattern for Example 1, (B) Data Pattern for Example 2, and (C) Data Pattern for Example 3.

weighting plays no role; however, for valid inferences the regression of Y on X needs to be correctly specified.

From this prediction perspective, $W = W(X)$ is simply a particular function of X, and can be considered a covariate; however, the balancing property of a propensity scores (Rosenbaum and Rubin, 1983) implies that, if the relationship between Y and W is correctly specified, relationships between Y and covariates other than W may be omitted or misspecified without biasing estimates of marginal parameters like means.

Example 2: Comparing two treatments with measured confounders X. Now suppose interest lies in the comparison of an outcome Y between two treatments defined by a binary treatment variable Z. Suppose that there are n units in the study, and units $i = 1, \ldots, n_0$ are assigned treatment $z_i = 0$ and units $i = n_0 + 1, \ldots, n_0 + n_1 = n$ are assigned treatment $z_i = 1$. Units are assumed randomly assigned with probabilities a function of known covariates X_i. We define $Y^{(t)}$ to be the outcome when assigned treatment $Z = t$. A frequentist MSM weights cases by the inverse of their allocation probability. From a Bayesian perspective, the task is to predict the values of $Y^{(t)}$ for the treatment t that was not assigned (see Figure 1B). As in Example 1, the approach is to build a regression model relating $Y^{(t)}$ to X in the treatment group $Z = t$, and use this model to predict the value of $Y^{(t)}$ in the treatment group $Z = 1 - t$. Weighting is replaced by prediction of the outcomes for treatments not assigned. As in Example 1, a model relating Y to the assignment propensity can be utilized instead of modeling all the covariates.

Example 3: A Bayesian MSM comparing two treatments at two time points. We now describe an MSM setting where intermediate variables serve as both mediators and confounders of a treatment effect. A classic example is the use of a surrogate marker for the outcome to make treatment decisions ("confounding by indication").

Suppose X_1 is a set of baseline covariates, Z_1 is a time 1 binary treatment indicator, assigned randomly with a probability depending only on X_1. A time 1 outcome X_2 is measured after the assignment of Z_1 , and then a time 2 binary treatment Z_2 is assigned with a probability that depends only on X_1, Z_1 and X_2. A time 2 outcome Y is then observed. Thus X_2 is both an outcome for the first treatment and the mediator for a second treatment. The causal inference concerns comparisons of the distributions of $Y^{(z_1,z_2)}$, the outcome when a subject is assigned Z_1 at the first time point and Z_2 at the second time point. The Bayesian analysis simulates predictions of the missing data, where the data are ordered as shown in Figure 1C. It has the following steps:

(a) Let $x_{i2}^{(z_1)}$ denote the intermediate value of X_2 when subject i is assigned to treatment z_1. Regress X_2 on X_1 and Z_1 and generate draws d of the missing values $x_{i2}^{(1,d)}$ for $i = 1, \ldots, n_0$ and $x_{i2}^{(0,d)}$ for $i = n_0 + 1, \ldots, n_0 + n_1$ from their respective predictive distributions.

(b) Regress Y on X_1, Z_1, X_2, Z_2 and use draws from the relevant regression model parameters to generate draws of the missing values $y_i^{(01,d)}, y_i^{(10,d)}, y_i^{(11,d)}$ for $i = 1, \ldots, n_{00}$, $y_i^{(00,d)}, y_i^{(10,d)}, y_i^{(11,d)}$ for $i = n_{00} + 1, \ldots, n_{00} + n_{01} = n_0$, $y_i^{(00,d)}, y_i^{(01,d)}, y_i^{(11,d)}$ for $i = n_0 + 1, \ldots, n_0 + n_{10}$, $y_i^{(00,d)}, y_i^{(01,d)}, y_i^{(10,d)}$ for $i = n_0 + 1, \ldots, n_0 + n_1$.

(c) Repeat (a) and (b) for a set of draws $d = 1, \ldots, D$, and simulate inferences for the causal parameters of interest from the observed or drawn values of $(y_i^{(00)}, y_i^{(01)}, y_i^{(10)}, y_i^{(11)}), i = 1, \ldots, n$.

No weighting is required, but as in previous examples, regressions needs to be correctly specified and estimable given the available data, as well as the standard sequential ran-

Table 1

Simulation study results under (a) correctly specified models ($\theta_{11} = -.187$, $\theta_{10} = -.068$, $\theta_{01} = -.095$); (b) misspecified second treatment model $Z_2 \mid X_1, X_2, Z_1$ ($\theta_{11} = -.079$, $\theta_{10} = -.025$, $\theta_{01} = -.035$); (c) misspecified intermediate variable model $X_2 \mid X_1, Z_1$ ($\theta_{11} = -.054$, $\theta_{10} = .020$, $\theta_{01} = -.095$)

	$\theta_{11} = E(Y(1,1) - Y(0,0))$			$\theta_{10} = E(Y(1,0) - Y(0,0))$			$\theta_{01} = E(Y(0,1) - Y(0,0))$		
Method	Bias	RMSE	95% Coverage	Bias	RMSE	95% Coverage	Bias	RMSE	95% Coverage
(a)									
Naive	.185	.193	3.2	.184	.192	5.4	−.042	.086	89.8
Weighted MSM	.000	.036	96.8	.000	.039	96.8	−.002	.060	94.5
Weighted SSMK	.001	.036	96.5	.001	.041	96.3	−.004	.061	94.1
Bayesian prediction	−.000	.030	96.0	.001	.024	98.1	−.001	.028	99.1
BP using MSM weights	.000	.036	91.8	.004	.034	95.6	−.004	.050	93.7
(b)									
Naive	.074	.078	8.7	.061	.065	22.0	−.026	.044	88.0
Weighted MSM	−.000	.021	96.2	.004	.024	92.5	.003	.020	95.2
Weighted SSMK	−.001	.022	95.8	.003	.024	92.9	.002	.020	94.1
Bayesian prediction	.000	.018	95.4	.001	.014	99.5	.000	.014	97.5
BP using MSM weights	−.000	.019	96.6	.001	.022	96.8	.003	.018	96.7
(c)									
Naive	.113	.120	26.1	.166	.170	0.0	−.033	.062	91.2
Weighted MSM	.002	.032	97.1	.000	.030	97.5	.001	.042	96.1
Weighted SSMK	.001	.032	97.2	.000	.030	97.1	−.001	.043	95.2
Bayesian prediction	−.005	.039	94.9	.008	.020	97.6	−.005	.032	98.5
BP using MSM weights	−.055	.066	70.1	.006	.024	95.4	−.048	.057	68.5

domization assumption for the outcome $Y^{(z_1,z_2)} \perp Z_1 \mid X_1$ and $Y^{(z_1,z_2)} \perp Z_2 \mid Z_1, X_1, X_2$. It also requires a similar sequential randomization assumption $X_2^{(z_1)} \perp Z_2 \mid Z_1, X_1$ for the intermediate variable X_2. Unlike the weighted MSM approach, correct models of the probability of intermediate treatment assignment are not required. An alternative approach, in the spirit of the propensity score discussion in Example 1, is to regress outcomes on weights $W(Z_{i1}, Z_{i2})$ associated with the assignment to each of the treatment conditions, namely $W^{-1}(Z_{i1}, Z_{i2}) = \frac{P(Z_{i2}|Z_{i1})P(Z_{i1})}{P(Z_{i2}|X_{i2},X_{i1},Z_{i1})P(Z_{i1}|X_{i1})}$.

We consider a simulation study to compare the Bayes prediction method with those discussed in SSMK. We generate data under a simple longitudinal model:

$$x_{i1} \sim N(0,1), \quad z_{i1} \mid x_{i1} \sim BIN(1, \text{expit}(-.1 + x_{i1}))$$

$$x_{i2} \mid x_{i1}, z_{i1} \sim N(x_{i1} - .75z_{i1}, 1/16),$$

$$z_{i2} \mid x_{i1}, x_{i2}, z_{i1} \sim BIN(1, \text{expit}(-.1 + (x_{i2} - x_{i1})(1 - z_{i1}))$$

$$y_i \mid x_{i1}, x_{i2}, z_{i1}, z_{i2} \sim BIN(1, \text{expit}(-.1 - z_{i2} + x_{i1} + x_{i2}))$$

where expit is the inverse of the logit function: expit(u)= $\frac{e^u}{1+e^u}$.

We consider three causal estimands of interest, corresponding to the differences in the expected probabilities of a "poor" outcome ($Y = 1$) when receiving no treatment compared with (a) both treatments (θ_{11}), (b) only the first treatment (θ_{10}), and (c) only the second treatment (θ_{01}), where $\theta_{kl} = E(Y^{(k,l)} - Y^{(0,0)})$. Thus treatment assignment at time 1, Z_1, is positively correlated with X_1, X_2 is a confounder/surrogate that mediates the effect of treatment assignment Z_1 on Z_2; assignment to treatment at time 1 is associated with smaller values of X_2, and assignment to treatment at time 2 is positively associated with increases in X_2 relative to X_1 in the absence of treatment at time 1 (suggesting treatment is "needed") and negatively associated with increases in X_2 relative to X_1 in the absence of treatment at time 1 (suggesting treatment is "failing"). The outcome Y is a function of only Z_2 given X_1 and X_2. Each simulation consisted of a sample of $n = 1000$ independently generated observations.

We fit five models to estimate the causal effect of Z_1, Z_2, and their interaction on Y: (1) a naive model that estimates a logistic regression for the observed Y on the observed X_1 and X_2; (2) a standard weighted MSM; (3) the weighted MSM method described in SSMK; (4) our Bayesian prediction (BP) model; and (5) a variation of the BP model that assumes the logit of the probability of the outcome is linearly related to $\log(W(Z_{i1}, Z_{i2}))$, omitting other covariates. Empirical Bias, RMSE, and nominal 95% coverage based on 1000 simulations are shown in Table 1.

The weighted MSM and Weighted SSMK methods perform very similarly in these simulations, perhaps in part because they are based on a large sample size. Table 1a) provides results under correctly specified models—the model on the log weights is approximately correctly specified. The naive model badly underestimates the marginal effect of treatment at time 2 and somewhat overestimates the effect of treatment at time 1 only, since subjects doing more poorly at time 1 are more likely to get treatment at time 2. All other approaches have minimal empirical bias, as expected given correct specification. The BP approach yields a reduction in RMSE of 17% for θ_{11}, 40% for θ_{10} and 60% for θ_{01} over the MSM approaches,

with the larger reductions to some degree associated with settings where weights inflate variance but have less of an impact in bias. Coverage for the MSM and BP approaches are general conservative. The BP approach using MSM weights is slightly biased, yielding increasing in RMSE and decreases in coverage, although RMSE is generally smaller than the MSM approaches, and coverage is correct to slightly undercovered.

Table 1b presents results when the second treatment assignment model is misspecified by omitting the interaction between X and Z_1—we generate $x_{i1} \sim N(3, 1)$ to enhance the effect of this misspecification, with the intercept for the outcome model set to -3. A modest degree of bias is introduced for the MSM and SSMK estimators, sufficient to induce undercoverage for θ_{10}. The BP approach yields minimal empirical bias, the best RMSE, and nominal or conservative coverage, as expected since the imputation models are correctly specified and misspecification of the weight model does not affect this method. Similar results are found for the BP using MSM weights, suggesting that misspecification of the weights themselves is less critical when they are used to model the outcome directly.

Table 1c shows results when the model for the weights is correctly specified, but the prediction model is misspecified by generating $x_{i2} \mid x_{i1}, z_{i1} \sim N(x_{i1} - .75z_{i1} + .5x_{i1}z_{i1}, 1/16)$, but ignoring the interaction in this model (again generating $x_{i1} \sim N(3, 1)$ and setting the outcome model intercept to -3 to enhance the effect of the misspecification). Predictably, both BP models are now biased for all the estimators, with large degrees of bias and undercoverage for the BP using MSM weights for θ_{11} and θ_{01}. Despite this, the direct BP prediction retains correct coverage and still has the smallest RMSE for θ_{10} and θ_{01}. The poorer behavior of the BP weight model even though the weight model is correctly specified is due to the fact that Y in no longer linear in $\log(W)$ as a result of the new model for X_2—a situation that could be remedied by modeling the relationship between Y and W more flexibly, for example by penalized spline regression (Zheng and Little, 2005).

In summary, we suggest that selection or allocation weights should be regarded as covariates in the Bayesian paradigm, and not used to weight for differential inclusion or differential allocation of units to treatments. We have presented an approach based entirely on prediction of missing variables for the sampled and/or nonsampled population—no hypothetical population is invoked.

References

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.

Rosenbaum, P. R. and Rubin D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Zheng, H. and Little, R. J. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics* **21**, 1–20.

Paul Gustafson*

Department of Statistics, University of British Columbia,
Vancouver, B.C., Canada V6T 1Z4

**email:* gustaf@stat.ubc.ca

1. Introduction

Saarela, Stephens, Moodie and Klein (SSMK hereafter) are to be congratulated for writing a really interesting paper. There has been considerable angst on how to reconcile Bayesianity with inverse probability of treatment (IPT) weighting. Fortunately, however, there have been recent insights. For instance, Wang, Parmigiani, and Dominici (2012)provide a model-selection sense in which a treatment-choice model matters for Bayesian inference, while Zigler et al. (2013) shed more light on the thorny issue of "feedback" from the outcome data to the treatment-choice model. And now SSMK may have moved the goalposts considerably. In the subtle longitudinal context, they provide a full and closely argued recipe for Bayesian inference based on modeling the treatment-choice mechanism. To help understand SSMK's method, here it is implemented in a simple example, and then compared to a different method.

2. Saturated Bayes Model

As a simple technical tool, a *Bayesian saturated binary regression model* (henceforth BSAT, for short) is defined as follows. For a binary response variable A and binary explanatory variables $B = (B_1, \ldots, B_p)$, a distinct outcome probability is assigned to each of the 2^p possible values of B, i.e., $\lambda_b = Pr(A = 1|B = b)$. As a conjugate prior, the elements of λ are taken as *iid*, with Unif$(0, 1)$ distributions. Hence a posteriori they are independently beta-distributed, and direct Monte Carlo sampling from the posterior distribution is trivial.

3. A Simple Worked Example

To bolster understanding of SSMK's proposal, consider the fictitious data in Table 1. These involve $n = 5000$ subjects, with only two timepoints ($m = 2$), two treatment alternatives ($r = 2$), a binary covariate, and a binary outcome. For ease of exposition, regard $Z_j = 1$ ($Z_j = 0$) as being "on" ("off") treat-

Table 1
An illustrative dataset on $n = 5000$ individuals with $m = 2$ timepoints. The time-varying treatment status Z_j, the time-varying covariate X_j, and the outcome Y are all binary.

X_1	Z_1	X_2	Z_2	$Y = 0$	$Y = 1$
0	0	0	0	1364	744
0	0	0	1	355	185
0	0	1	0	371	335
0	0	1	1	104	88
0	1	0	0	8	5
0	1	0	1	91	45
0	1	1	0	1	2
0	1	1	1	25	11
1	0	0	0	18	13
1	0	0	1	4	3
1	0	1	0	279	389
1	0	1	1	71	100
1	1	0	0	2	1
1	1	0	1	11	11
1	1	1	0	19	19
1	1	1	1	141	185

ment at the j-th timepoint, while $Y = 1$ is a harmful outcome, such as having reached irreversible disease progression by the end of the study. As per SSMK, consider the variables as unfolding in the temporal order (X_1, Z_1, X_2, Z_2, Y).

In such a small-scale problem there is no need to make parametric modeling assumptions. Each treatment pattern $\tilde{a} = (v, w)$ can have a distinct counterfactual mean outcome θ_{vw}. Focussing in on the "always-treat" pattern, $\tilde{a} = (1, 1)$, the subset of the data pertaining to those treated at both timepoints is summarized in Table 2. Taking these data at face value, without any regard to confounding, yields an estimated outcome probability for the always-take regime to be $252/(252 + 268) = 0.485$.

Upon fitting BSAT models for (Z_1), $(Z_2|Z_1)$, $(Z_1|X_1)$, and $(Z_2|X_1, Z_1, X_2)$, and then exactly following SSMK's prescription (as per their Supplementary Appendix B), a weight is estimated for each (X, Z) pattern. For $Z = (1, 1)$ these are reported in Table 2, and they yield the weighted (X, Y) data table given in the rightmost columns of the table, i.e., raw cell counts in each row get multiplied by the corresponding weight. The weighted table then gives an estimated outcome probability of $200.59/(319.52 + 200.59) = 0.386$ for the counterfactual world in which everyone is fully treated. The adjustment for time-varying confounding has quite markedly changed the face-value impression of the data.

Table 2
Summary of the 520 individuals treated at both timepoints. Both the raw and reweighted X by Y data tables are given, with the latter obtained from the former by row multiplication by the estimated weight $\widehat{w}_i$ for an individual with this X pattern and $Z = (1, 1)$.

X_1	X_2	$Y = 0$	$Y = 1$	$\widehat{w}$	$Y = 0$	$Y = 1$
0	0	91	45	2.26	205.28	101.51
0	1	25	11	2.27	56.71	24.95
1	0	11	11	0.40	4.36	4.36
1	1	141	185	0.38	53.17	69.76
Total		268	252		319.52	200.59

SSMK's method does not directly use the weighted columns of Table 2, yet reweighting rows of X by Y data tables, for fixed Z, is central. Particularly, the proposal is to Bayesian-bootstrap first, reweight second. Each individual's contribution of exactly one datum is replaced with a contribution of roughly one datum, via the Dirichlet sampling scheme. This gives a "noised up" X by Y data table for fixed Z, which is then reweighted to yield a point estimate of the corresponding θ_{vw}. Then the ensemble of such point estimates arising from repeated Bayesian bootstrapping is taken as the posterior distribution of θ_{vw}. Doing this, we obtain a posterior mean (SD) of 0.386 (0.027) for θ_{11}. Not surprisingly given the nature of bootstrapping, the posterior mean is effectively the same as the point estimate from the reweighted portion of Table 2. Of course, interest typically lies in comparing different regimes. For $\theta_{11} - \theta_{00}$, the counterfactual risk difference between fully treating and fully not treating, we get a posterior mean (SD) of -0.047 (0.029).

4. Using the g-Formula

Stepping back, IPT weighting is but one strategy for such problems. The *g-formula* approach pioneered by Robins (1986) is another. There are various accounts of this method; see an Appendix to Taubman et al. (2009) for an accessible explanation. The core idea is to probabilistically express the time-evolution of all variables, including both the treatment choice arising under a specific intervention and the treatment choice arising in the absence of any intervention. Assumptions about no unmeasured confounding then equate to assumptions about this joint distribution. In the special case of a single timepoint, the g-formula reduces to the well-known epidemiological procedure of standardization (Snowden, Rose, and Mortimer, 2011).

For the problem at hand, the g-formula specializes as follows. The outcome probability in the counterfactual world where all are fully treated reduces to

$$\theta_{11} = \sum_{x_1=0}^{1}\sum_{x_2=0}^{1} [Pr(X_1 = x_1)Pr(X_2 = x_2|X_1 = x_1, Z_1 = 1) \times Pr\{Y = 1|X = (x_1, x_2), Z = (1, 1)\}],$$

with an analogous expression obtained for θ_{00}. Thus BSAT models can be fit for (X_1), $(X_2|X_1, Z_1)$, and $(Y|X_1, Z_1, X_2, Z_2)$, requiring 1, 4, and 16 parameters, respectively. Let γ_1, γ_2, and γ_3 be the three parameter vectors, with $\gamma = (\gamma_1, \gamma_2, \gamma_3)$. It is immediate that if the γ_i's are judged to be a priori independent of one another, then they will be a posteriori independent also. So direct *iid* Monte Carlo sampling of the posterior on γ is trivial. Moreover, since each θ_{vw} is a deterministic function of γ, we obtain the posterior for θ with no additional fuss or cost. Applying this to the Table 1 data yields a posterior mean (SD) for $\theta_{11} - \theta_{00}$ of -0.043 (0.028). The Bayesian g-formula (BGF) and SSMK answers are essentially the same.

5. Simulation

Before investigating further, we disclose the origins of the Table 1 data. Each individual's data was simulated (forward in time) according to:

$$
\begin{aligned}
X_1 &\sim \text{Bern}(0.25) \\
Z_1 &\sim \text{Bern}(0.05 + \kappa_1 X_1) \\
X_2 &\sim \begin{cases} \text{Bern}(0.25 - \kappa_2 Z_1) & \text{if } X_1 = 0 \\ \text{Bern}(0.95) & \text{if } X_1 = 1 \end{cases} \\
Z_2 &\sim \begin{cases} \text{Bern}(0.1 + \kappa_1 X_2) & \text{if } Z_1 = 0 \\ \text{Bern}(0.9) & \text{if } Z_1 = 1 \end{cases} \\
Y &\sim \text{MaxBern}(0.2 - \kappa_3 Z_1 + \kappa_4 X_1 \ , \ 0.2 - \kappa_3 Z_2 + \kappa_4 X_2),
\end{aligned}
$$

where MaxBern(a, b) is taken to be the distribution of the maximum of independent Bern(a) and Bern(b) random variables. Thus, we mimic the realistic situation of an irreversible disease outcome that could be reached after the first or second timepoint.

The key drivers of the scenario are set as $\kappa_1 = 0.25$, $\kappa_2 = 0.075$, $\kappa_3 = 0.03$, and $\kappa_4 = 0.15$. By dint of all these being positive, the scenario has some typical features. Think of $X_j = 1$ as some manifestation of "being sicker" at the j-th timepoint, as meshes with $\kappa_4 > 0$. Via $\kappa_1 > 0$, those who are sicker are more likely to start treatment, a hallmark of "confounding by indication." Also, positive κ_2 and κ_3 reflect a dual benefit of treatment. There is a direct effect (controlled by κ_3), but also an indirect effect (controlled by κ_2), whereby treatment reduces the chance of the undesirable transition from $X_1 = 0$ to $X_2 = 1$.

Figure 1 gives results for 50 datasets simulated as described. The previously seen agreement in estimating $\theta_{11} - \theta_{00}$ is no fluke; the SSMK and BGF posterior means are always essentially the same. The corresponding posterior standard deviations agree less closely, though both exhibit very modest variation across repeated sampling. Moreover, in an extended simulation of 1000 datasets, the empirical coverage of 95% equal-tailed credible intervals are 95.6% and 95.7% for SSMK and BGF, respectively, with a discordance rate (one interval covers but the other does not) of only 0.9%. Taking stock then, the two methods start with different premises, and require modeling different parts of the joint distribution of observables. Yet here they give essentially the same estimate and about the same indication of uncertainty.

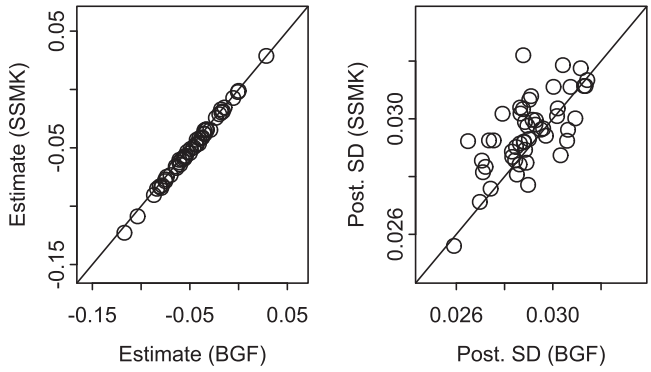

Figure 1. Comparing BGF and SSMK inferences on 50 simulated datasets. The left panel compares posterior means and the right panel compares posterior standard deviations.

6. Discussion

It is comforting to see two seemingly disparate statistical methods, SSMK and BGF, yielding the same answer. In contrasting the methods, however, BGF does have a simple elegance. Models for some conditional distributions must be specified, along with concomitant priors. Then the Bayesian crank is turned, and the target parameter is a function of the unknown parameters. The need for bootstrapping is obviated, and special arguments to support the method's validity are not required.

References

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**, 1393–1512.

Snowden, J. M., Rose, S., and Mortimer, K. M. (2011). Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology* **173**, 721–730.

Taubman, S. L., Robins, J. M., Mittleman, M. A., and Hernán, M. A. (2009). Intervening on risk factors for coronary heart disease: An application of the parametric g-formula. *International Journal of Epidemiology* **38**, 1599–1611.

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

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.

Alessandra Mattei* and Fabrizia Mealli**

Department of Statistics, Computer Science, Applications, University of Florence, Florence, Italy

**email:* mattei@disia.unifi.it

***email:* mealli@disia.unifi.it

1. Introduction

In the ambitious article by Saarela, Stephens, Moodie and Klein (hereafter SSMK), the authors propose a Bayesian perspective on causal inference for longitudinal treatments under sequential ignorability (SI) assumptions using marginal structural models (MSMs) and inverse-probability-of treatment weighted (IPTW) methods. Their approach can be viewed as

blending two strands of the literature in causal inference: (1) MSMs and IPTW estimators are often applied to consistently estimate causal effects, although not in a Bayesian perspective; (2) recent work has focused on Bayesian inference for causal effects in observational studies using propensity score methods (see references in the SSMK's article). To the best of our knowledge, the only attempt to develop a fully Bayesian approach to causal inference for sequential treatments is in Zajonc (2012), who does not use propensity score methods.

Although we find the authors' contribution of some interest, the reading of the article stimulated some reactions. The points we wish to make and discuss include (1) the importance to clearly define the causal estimands of interest; (2) the role of the weights in a fully Bayesian approach; and (3) the role of the assignment mechanism to design observational studies.

As we better explain in the sequel, our main disappointment with the article arises because the authors provide several technical details on the Bayesian interpretation of the weights used by IPTW estimators, neglecting substantive discussion on the assumptions underlying the definition of the weights and of a MSM. Specifically, the authors implicitly assume that treatment assignment model and weights are correctly specified, and that weights are successful in creating a pseudo-population where the distributions of relevant confounders are overlapping and well balanced across units exposed to alternative treatment sequences. We argue that assessing the degree of overlap in the confounders' distributions and the balancing property of weights is crucial and should be the only guidance in the specification of the weights. This point relates also to the paramount importance of carefully designing an observational study, which is essential for drawing credible objective causal inference, especially in the presence of sequential treatments.

2. Causal Estimands

In causal inference problems, it is crucial to start by precisely defining the causal quantities (causal estimands) we want to draw inference on. In the SSMK's article, the definition of the causal estimand of interest is a bit nebulous and the presentation is sometimes unclear and ambiguous. SSMK focus on inferential issues omitting details on the link between potential outcomes and observed outcomes in a MSM framework, and thus making it difficult to understand when they are working with the parameters of the MSM, and thus with potential outcomes, and when, instead, they are considering the parameters of an associational model for the observed outcomes.

In order to shed light on these key issues, we find it useful to reformulate the basic setup according to our understanding and our view. We will follow the notation used by the authors as much as possible.

Consider a sample of n units, indexed by $i = 1, \ldots, n$. In each time period, indexed by $j = 1, \ldots, m$, units can be potentially assigned to r alternative treatments, $a_j \in \{w_1, \ldots, w_r\}$. Let $\tilde{a} \equiv (a_1, \ldots, a_m)$ denote one of the r^m treatment sequences, which a unit can be potentially assigned to. The objective is to assess the effect of different treatment sequences on some final outcome, y, measured after the assignment of the last treatment, a_m. Under the Stable Unit Treatment Value Assumption (SUTVA, Rubin, 1980), for each unit there are r^m associated potential outcomes at a future point in time after the last treatment, but at most one of which can be observed. Let $y_{\tilde{a}i}$ be the value of y if unit i is assigned treatment sequence $\tilde{a}$, and let $\mathbf{y}_{\tilde{a}i}$ denote the set of all the r^m potential outcomes for unit i. A causal effect of two alternative treatment sequences, $\tilde{a}$ and $\tilde{a}'$, on a unit is defined to be a comparison of potential outcomes, for instance, their difference $y_{\tilde{a}i} - y_{\tilde{a}'i}$.

Let z_{ij} denote which treatment unit i received at time j, and let $\tilde{z}_i \equiv (z_{i1}, \ldots, z_{im})$ denote the observed history of treatment assignments. Let $y_i = \sum_{\tilde{a}} \mathbf{1}_{\{\tilde{z}_i = \tilde{a}\}} y_{\tilde{a}i}$ be the observed outcome, where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. For each units, we also observe the set of vectors $\tilde{x}_i \equiv (x_{i1}, \ldots, x_{im})$, including both baseline covariates, measured prior to the onset of treatment, and time-varying covariates (intermediate outcomes) recorded prior to each treatment assignment. Finally, let $\tilde{z}_{i(j)} \equiv (z_{i1}, \ldots, z_{ij})$ and $\tilde{x}_{ij} \equiv (x_{i1}, \ldots, x_{ij})$ denote the observed treatment vector and the observed vectors of covariates up to time j, $j < m$.

This is the basic setup, which defines the primitives for causal inference and the observed data in longitudinal studies. The next step is to introduce some structural and, eventually, parametric assumptions that allow us to draw inference on the causal estimands of interest. MSMs describe the distribution of potential outcomes, or some of their characteristics, in terms of parameters of specific models: $y_{\tilde{a}i} \sim p(\cdot; \theta)$, where $p(\cdot; \theta)$ denote a probability density/mass function with parameter vector θ. For instance, if the outcome variable is binary, and the treatment assigned in each period is binary ($r = 2$, $a_j \in \{0, 1\}$, $j = 1, \ldots, m$), a MSM is usually specified as a linear logistic model: $\text{logit}\,(\pi_{\tilde{a}i}) \equiv \text{logit}\,(Pr(y_{\tilde{a}i} = 1)) = \theta_1 + \theta_2 \sum_{j=1}^{m} a_j$. In this perspective, causal estimands can be usually expressed in terms of the parameters, θ, of the MSM, which therefore represent comparisons of potential outcomes. A MSM may impose restrictive constraints on causal effects. For instance, in the previous example, the logistic MSM implies that causal effects of alternative treatment sequences are additive on logit scale, so that $e^{\theta_2} = [\pi_{\tilde{a}i} \cdot (1 - \pi_{\tilde{a}i})^{-1}]/[\pi_{\tilde{a}'i} \cdot (1 - \pi_{\tilde{a}'i})^{-1}]$ represents the causal odds ratio for the comparison of all treatment sequences $\tilde{a}$ and $\tilde{a}'$ such that $\sum_{j=1}^{m} a_j - \sum_{j=1}^{m} a'_j = 1$ (e.g., $\tilde{a} = (1, 0, \ldots, 0)$ and $\tilde{a}' = (0, 0, \ldots, 0)$). Thus, in a MSM approach focus is on identifying and estimating the parameter vector θ. We believe that explicitly motivating the interest in the parameters θ is crucial to avoid misunderstandings. SSMK focus on deriving the posterior distribution of θ, but do not clearly explain the role of the parameter θ in the causal problem.

Another related issue that we believe the authors should better discuss concerns the role of the assignment mechanism. The critical assumption invoked in the article is SI, which the authors refer to as *no unmeasured confounding/sequential randomization* condition: $p(\tilde{z}_i | \mathbf{y}_{\tilde{a}i}, \tilde{x}_i) = p(z_{i1} | x_{i1}) \prod_{j=1}^{m} p(z_{ij} | \tilde{z}_{i(j-1)}, \tilde{x}_{ij})$. An overlap assumption is also invoked: $p(z_{ij} = a_j | \tilde{z}_{i(j-1)}, \tilde{x}_{ij}) > 0$ for all i, j, and $\tilde{a}$.

It is well known that the parameters of MSMs, θ, differ from the parameters of associational models for the observed outcome, say $p(y_i | \tilde{z}_i, \theta^*)$. However, under SI, the causal parameter θ of a MSM can be consistently estimated using an IPTW estimator, that is, fitting the corresponding associational model for the observed outcome, $p(y_i | \tilde{z}_i, \theta^*)$, with the stabilized weights $w_i = [p(z_{i1}) \prod_{j=2}^{m} p(z_{ij} | \tilde{z}_{i(j-1)})]/$

$[p(z_{i1}|x_{i1}) \prod_{j=2}^{m} p(z_{ij}|\tilde{z}_{i(j-1)}, \tilde{x}_{ij})]$. The authors do not clearly distinguish between parameters of associational models, θ^*, and parameters of MSMs, θ, and we believe that the lack of clarity on this key point may raise misunderstandings.

SSMK put a great deal of effort to prove that IPTW estimation can be derived through a Bayes decision rule, defining a conceptual population where a longitudinal randomized experiment has been conducted. We appreciate the idea, which might also thrill Bayesians.

However, weighting by the inverse of selection probabilities is a well-known approach, not only in causal inference, but in statistics in general, which dates back to Horvitz and Thompson in 1952. In causal inference, the interpretation of the weights as a tool to construct a pseudo-population, which resembles a "target" population where there is no imbalance in the observed variables is a weighting approach's own feature; it is not related to any specific inferential approach. According to us, the critical issue underlying a weighting approach is not the interpretation of the weights, but rather the specification of the model used to estimate the weights, just as in observational studies with single-time point treatments focus is on the specification of the propensity score, in order to obtain estimates of the propensity score that balance the pretreatment covariates between treatment groups, also accounting for the degree of overlap in the covariate distributions.

The problem to assess the overlap condition (positivity assumption) and to find a specification of treatment assignment model leading to weights that satisfy the balancing property is even more dramatic in longitudinal observational studies, where we need to control not only for pretreatment baseline covariates, but also for intermediate variables, which are simultaneously posttreatment outcomes and pretreatment confounders.

3. The Role of the Weights in a Fully Bayesian Approach

Although the article promises to develop a 'fully Bayesian procedure for estimating parameters of MSMs using IPT weighting,' the reading of the article may leave a nasty taste in a Bayesian's mouth. The approach proposed by SSMK is far from standard Bayesian approaches, and in our view, it is not actually *fully* Bayesian.

From a Bayesian perspective, all aspects of the observable data must be modeled, and the target of inference is the joint posterior distribution of all model parameters including the parameters of the outcome model and the parameters of the treatment assignment model. According to this view, Bayesian inference for time-varying treatments follows by specifying a joint model for intermediate confounders, final potential outcomes and treatment assignment probabilities (and possibly even for baseline covariates). Assumptions may help simplifying the model. Under SI, the treatment assignment model does not depend on missing intermediate and final outcomes. Therefore, a fully Bayesian approach for longitudinal treatment under SI requires only a model for the joint distribution of intermediate confounders and final potential outcomes (and baseline covariates), at least if the parameters of the treatment assignment model and the parameters of the outcome model are a priori independent (see also Gustafson, 2012). Posterior distributions of causal effects are derived by marginalizing over appropriate covariates distributions, as in g-estimation methods. Thus, in a fully Bayesian perspective weights can be ignored, because they do not enter the posterior distribution of the causal parameters.

A fully Bayesian approach may end up with a very complex model raising inferential challenges. However modern methods of computational statistics can make inference relatively straightforward, as shown in Zajonc (2012) and suggested by Gustafson (2012).

4. Designing a Longitudinal Observational Study

The importance of carefully designing observational studies with single-time, often binary, treatments is well-known in the causal inference literature (e.g., Rubin, 2008). We argue that carefully designing a longitudinal observational study is also of paramount importance.

Under SI, the core of the design phase of a study with sequential treatments is to create a subpopulation of units exposed to alternative treatment sequences, where the distributions of the confounders, including both pretreatment baseline covariates and intermediate (time-variant) confounder variables, are overlapping and well balanced across units exposed to alternative treatment sequences.

SSMK repeatedly stress that IPTW methods aim at constructing a pseudo-population without imbalances in observed variables, formally defining the weights in terms of treatment assignment probabilities. According to us, the specification of treatment assignment probabilities, which need to be estimated from the data, deserves special attention, just as the specification of the propensity score in non-longitudinal settings. Indeed MSMs can be sensitive to misspecification of treatment assignment model (e.g., Imai and Ratkovic, in press).

We argue that the specification of the treatment assignment model should be judged aiming at obtaining weights that balance the distributions of baseline and time-varying covariates across all appropriate sub-populations (e.g., Imai and Ratkovic, in press).

The performance of IPTW estimators also depends on the positivity (overlap) assumption. Assessing overlap in covariate distributions is crucial because, even if SI holds, there may be regions of the covariate space with relatively few units exposed to specific treatment sequences. In such a case, some weights can be extremely large making some units particularly influential, and thus making inferences on causal effects less precise.

The authors completely neglect how the problem of assessing the degree of overlap in the confounder's distributions can be addressed in a longitudinal setting and provide no discussion on the specification of the weights. They simply invoke the positivity assumption and implicitly assume that specification of the treatment assignment model is correct. Conversely, we argue that a critical specification of the weights is essential, as also pointed out by a recent strand of the literature, which is moving toward a better thought out construction of weights in longitudinal observation studies, proposing methods to assess overlap in the covariate distributions and evaluate the balancing property of the weights (Achy-Brou,

Frangakis, and Griswold, 2010; Platt, Delaney, and Suissa, 2012; Imai and Ratkovic, in press).

Once a sub-population where there is overlap in the distributions of the observed variables and there is no imbalance in the observed variables has been constructed (e.g., by trimming and/or matching using the treatment assignment probabilities), one can move to the analysis phase. Given a good design phase, in the analysis phase one can use any procedures for estimating causal effects, including IPWT estimators and Bayesian imputation methods.

SSMK emphasize that MSMs, coupled with weighting, do not require to explicitly model intermediate variables and relationships. We do not question the benefits that MSMs may offer, but we believe that the authors should stress that the advantages of MSMs are derived from the underlying assumptions of SI, positivity and correct specification of both the treatment assignment model and the MSM. Correct MSM specification is another critical assumption entering the analysis phase. We can understand that misspecification in the MSM is a topic that goes beyond the aim of the article, but feel the authors should have been clear about the critical assumptions and provided some discussion on their plausibility, at least in the empirical study.

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.

Gustafson, P. (2012). Double-robust estimators: Slightly more Bayesian than meets the eye? *The International Journal of Biostatistics* **8**, 1–15.

Imai, K. and Ratkovic, M. (in press). Robust estimation of inverse probability weights for marginal structural models. *Journal of the American Statistical Association. in press.*

Platt, R. W., Delaney, J. A., and Suissa, S. (2012). The positivity assumption and marginal structural models: The example of warfarin use and risk of bleeding. *European Journal of Epidemiology* **27**, 77–183.

Rubin, D. B. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *Journal of the American Statistical Association* **75**, 591–593.

Rubin, D. B. (2008). For objective causal inference, design trump analysis. *The Annals of Applied Statistics* **2**, 808–840.

Zajonc, T. (2012). Bayesian inference for dynamic treatment regimes: Mobility, equity, and efficiency in student tracking. *Journal of the American Statistical Association* **107**, 80–92.

James M. Robins, Miguel A. Hernán*

Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A
Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A
**email:* miguel_hernan@post.harvard.edu

and

Larry Wasserman

Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A

Saarela et al. are concerned with integrating propensity scores into a Bayesian framework. Some of us have previously written (Robins and Ritov, 1997; Robins and Wasserman, 2000; http://normaldeviate.wordpress.com/2012/08/28/robins-and-wasserman-respond-to-a-nobel-prize-winner/; posted 28 Aug 2012, accessed 1 Oct 2014) about this topic, every time making much the same argument. Here, we present a simplified version that captures the main points.

A Simple Setting

Though our argument applies to the complex observational data considered by Saarela et al., it is easier to understand it in the simpler setting of a double-blind, placebo-controlled randomized clinical trial of a non-time-varying treatment and under complete compliance. In the spirit of the authors, we assume the trial subjects are representative of a much larger population and the trial results will guide treatment decisions in the population.

Let $\mathbf{V} = \{Z_i, X_i; Y_i; i = 1, \ldots, n\}$ denote the data on the n trial subjects, where Z_i is the binary treatment arm indicator, Y_i is the binary outcome, and X_i is a high-dimensional vector of baseline covariates. The randomization probabilities $\mathrm{pr}[Z = 1|X]$ are chosen by a randomizer. By de Finetti's theorem (e.g., Bernardo and Smith, 1994), a Bayesian can write the marginal density $p(\mathbf{V})$ of V

$$p(\mathbf{V}) = \int_{\phi,\gamma} p(\mathbf{Z}, \mathbf{X}, \mathbf{Y}; \phi, \gamma)\, p(\phi, \gamma)\, \mathrm{d}\mu(\phi)\, \mathrm{d}\mu(\gamma),$$

$$p(\mathbf{Z}, \mathbf{X}, \mathbf{Y}; \phi, \gamma) = \mathcal{L}_1(\phi)\, \mathcal{L}_2(\gamma),$$

$$\mathcal{L}_1(\phi) = \prod_{i=1}^{n} L_{1i}(\phi), \quad \mathcal{L}_2(\gamma) = \prod_{i=1}^{n} L_{2i}(\gamma),$$

where $L_1(\phi) = f(Y|Z, X; \phi_1)\, f(X; \phi_2)$ and $L_2(\gamma) = f(Z|X, \gamma)$. We have already integrated out the authors' unmeasured frailty U.

The propensity score $e(X;\gamma^\dagger) = \text{pr}[Z=1|X;\gamma^\dagger]$ is known to the randomizer by design, but let us provisionally assume that our Bayesian does not know it so he treats γ as random. [We assume there exist true values $\left(\phi^\dagger, \gamma^\dagger\right)$ of (ϕ, γ) but, even if not, our argument, slightly modified, is still valid].

Like the authors, we take our goal to be the estimation of the counterfactual probabilities $\theta^\dagger = \left(\theta_0^\dagger, \theta_1^\dagger\right)$, where $\theta_z^\dagger = \text{pr}(Y_z = 1)$, and Y_z is a subject's counterfactual response under treatment level z. Randomization implies that $\theta_z^\dagger$ is identified and equals

$$\theta_z^\dagger = \int \text{pr}\left[Y=1|Z=z, X=x;\phi_1^\dagger\right] f\left(x;\phi_2^\dagger\right) \text{d}x$$

Why Bayesian Inference Must Ignore the Propensity Score

Bayesian logic is rigidly defined: given a likelihood and a prior, one turns the Bayesian crank to obtain a posterior. There is no wiggle room. A fact concisely summarized in the slogan "There is no Bayes but Bayes." Because the parameter θ of interest is a functional of the parameters ϕ, the posterior for θ is completely determined by the posterior of ϕ. If ϕ and γ are a priori independent, the posterior of ϕ is obtained from the $\mathcal{L}_1(\phi)$ factor of the observed likelihood and the prior $p(\phi)$ for ϕ.

Therefore, Bayesian inference concerning θ cannot be a function of the propensity score $e(X;\gamma^\dagger)$ because the Bayesian's posterior for ϕ—and thus for θ—does not depend on γ. Saarela et al. assume ϕ and γ are a priori independent and yet argue that inverse probability weighting by a function of the propensity score $e(X;\gamma^\dagger)$ can be given a Bayesian interpretation. In light of the above, their arguments cannot be valid.

Why Propensity Scores Should Not be Ignored

Why do the authors, as Bayesians, work so hard to include propensity scores in their inference when, according to Bayes, they are irrelevant? Our guess is that the authors recognize that an analysis—Bayesian or otherwise—that ignores a known propensity score can go seriously wrong because one's prior knowledge of $\text{pr}[Y=1|Z=z, X]$ is meager when X is high-dimensional.

Specifically, consider any estimator $\widehat{\theta}$ of $\theta^\dagger$ that does not depend on the known propensity score. Robins and Ritov (1997) prove that $\widehat{\theta}$ cannot be uniformly consistent for $\theta^\dagger$ over the large infinite dimensional model $\mathcal{M}$ that includes any laws $b_z(X) = \text{pr}[Y=1|Z=z, X]$, any density $f(x)$ for X, and any propensity function $e(X) = \text{pr}[Z=1|X]$ bounded away from 0 and 1. The practical implication of this theorem is that, whenever $e(X;\gamma^\dagger)$ is a complex function of our high dimensional X and the (infinite-dimensional) parameters γ and ϕ are a priori independent, the posterior for θ will fail to concentrate around the true value of $\theta^\dagger$ as n goes to infinity because any model we specify for $f(Y|Z, X;\phi_1)$ is almost certainly incorrect (imposing smoothness will not really help). This practical implication is obvious; the Robins and Ritov theorem serves as a mathematical formalization.

In contrast, estimators that use the known randomization probabilities, like the Horvitz-Thompson (1952) estimator of $\theta_z^\dagger$, can be uniformly $n^{1/2}$-consistent over $\mathcal{M}$. The deficiencies of the Horvitz-Thompson estimator—it may exceed 1, it ignores data on X except for the one-dimensional summary $e(X;\gamma^\dagger)$, and it can be very inefficient—can be remedied by using an improved version: the so-called *locally semiparametric efficient regression estimator* (Scharfstein, Rotnitzky, Robins 1999). In observational studies, this estimator is doubly robust when the unknown $e(X;\gamma^\dagger)$ is replaced by an estimate. More efficient doubly robust estimators are reviewed by Rotnitzky et al. (2012).

When the Priors are Dependent

Our argument relies on the authors' assumption that ϕ and γ are a priori independent. This assumption is often reasonable, as shown in the Appendix. However, when ϕ and γ are a priori dependent—which implies that the posterior for θ will depend on the propensity score $e(X;\gamma)$—two new issues arise.

First, in observational studies with $\gamma^\dagger$ unknown, the posterior for γ will depend on the data through the ϕ part of the likelihood. The authors find this troubling since this procedure fails to "retain the balancing property of propensity scores." But again true Bayesians cannot have it both ways. The parameters ϕ and γ are either a priori independent or they are not. If one wants to use dependent priors to make the posterior for θ to depend on the propensity score, then one must accept that the posterior for the propensity score will depend on the ϕ part of the likelihood.

The above is not only a philosophical issue concerning schools of inference. It implies that true Bayesian inference based on finite-dimensional working models will generally fail to be doubly robust since misspecification of either the outcome or propensity model will bleed into the estimation of the parameters of the other correct model. As the authors discuss in their supplemental material, this lack of double robustness confronted both McCandless et al. (2010) and Zigler et al. (2013) who proposed approaches to prevent the bleeding. But, as useful as the approaches may be, they cannot be truly Bayesian.

Second, even in a randomized trial with known propensity score, simply making ϕ and γ dependent a priori does not imply that the posterior for θ will concentrate around the truth. The dependent prior still has to be carefully engineered for that to happen. As an example we can construct a locally semiparametric efficient Bayes estimator $\widehat{\theta}_{\text{Bayes}}$ as follows. We assume that, conditional on the known $\gamma^\dagger$ and k given functions $w_{m,z}(x)$, $\text{pr}(Y=1|Z=z, X=x;\phi_{1,z})$, $\phi_{1,z} = (\eta_{1,z}, \ldots, \eta_{k,z})$ is a finite-dimensional parametric function $\text{expit}\left\{\sum_{m=1}^k \eta_{m,z} w_{m,z}(x)\right\}$ with $w_{k,z}(x) = 1/\text{pr}(Z=z|X=x;\gamma^\dagger)$. Then, if we put smooth or non-informative priors over the parameters $\phi_{1z} = (\eta_{1,z}, \ldots, \eta_{k,z})$, the Bayes estimator $\widehat{\theta}_{\text{Bayes}}$ will be asymptotically equivalent to the frequentist locally semiparametric efficient estimator cited earlier and thus be $n^{1/2}$-consistent. Thus, by using carefully tuned dependent priors, we have obtained a Bayes estimator that has good frequentist behavior by mimicking a locally semiparametric efficient frequentist estimator.

But this is a Pyrrhic victory. If we need to engineer the dependent prior just to mimic a frequentist answer, is it really Bayesian inference? We call Bayesian inference which is care-

fully manipulated to force an answer with good frequentist behavior, *frequentist pursuit*. There is nothing wrong with it. But if you want to be Bayesian, then accept that, in this example, your posterior will fail to concentrate around the true value.

Conclusion

Our arguments above may have left readers thinking "why bother? If you want good frequentist properties, just use a frequentist estimator rather than embarking on a frequentist pursuit." Indeed, it might appear that we are arguing that the Bayesian machinery should be reserved for implementing subjective Bayes inference that maps prior beliefs to posterior beliefs via the likelihood function, without regard for the frequentist properties of the resulting estimators. While we do believe that investigation of this mapping through Bayesian sensitivity analysis and/or robust Bayes is important and extremely useful, we also believe that the Bayesian approach can play other important roles, even when one is interested in good frequentist properties. We consider three cases.

First, Bayesian logic and machinery may sometimes lead to procedures with provably better frequentist operating characteristics than their current competitors, even asymptotically. An example is the conditional predictive and partial posterior predictive p-values of Bayarri and Berger (2000).

Second, when modeling complex phenomena (particularly in small and moderate samples), there may be Bayesian approaches that are rather straightforward to motivate and implement even when there is no good frequentist alternative, so the Bayes estimator is the best, or perhaps the only, frequentist game in town.

Third, to improve decision making under uncertainty, one can adopt a Bayes-frequentist compromise (Robins 2004, Sec 5.2) that combines honest subjective Bayesian inference with good frequentist behavior even when, as above, the model is so large and the likelihood function so complex that standard (uncompromised) Bayes procedures have poor frequentist performance. It follows immediately from our earlier arguments that such a compromise requires that our subjective Bayesian decision maker is only allowed to observe a specified vector function of X (depending on $e(X;\gamma^{\dagger})$) but not X itself. In this way one can circumvent the problem referred to by Robert (http://xianblog.wordpress.com/2013/01/17/robbins-and-wasserman; posted 17 Jan 2013, accessed 01 Oct 2014) as the *curse of marginalization*: "the classical Bayesian approach is an holistic system that cannot remove information to process a subset of the original problem."

Acknowledgement

This work was partly funded by NIH grants P01 CA134294 and R01 AI102634.

References

Bayarri, M. J. and Berger, J. O. (2000). P-values for composite null models. *Journal of the American Statistical Association* **95**, 1127–1142. Rejoinder, pp 1168–1170.

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

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

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**, 16.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In D. Y. Lin, P. Heagerty (eds). Proceedings of the Second Seattle Symposium on Biostatistics. New York: Springer.

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

Robins, J. M. and Wasserman, L. (2000). Conditioning, likelihood, and coherence: a review of some foundational concepts. *Journal of the American Statistical Association* **95**, 1340–1346.

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99**, 439–456.

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**, 1096–1120.

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.

Appendix

Example of A Priori Independence of the Propensity Score

Suppose a health insurance company needs to estimate the fraction θ of its patient population that will have a myocardial infarction (MI, $Y = 1$) in the next year, so as to determine the need for cardiac unit beds. They have 300 potential risk factors $X = (X_1, \ldots, X_{300})$ measured on each member. A general epidemiologist had earlier studied risk factors for MI by following 5000 patients for a year. Because MI was a rare event, he oversampled subjects whose X, in his opinion, indicated a higher conditional probability $b(x) = \mathrm{E}[Y|X = x]$ of $Y = 1$. Hence, with Z the inclusion indicator, the sampling fraction $e(x) = \mathrm{pr}(Z = 1|X = x)$ was a known but complex function.

The world's leading heart expert, our Bayesian, was hired to estimate $\theta = \int b(x)\, p(x)\, \mathrm{d}x$, where $p(x)$ is the marginal density of x, based on the study data $(\mathbf{X}, \mathbf{Z}, \mathbf{ZY})$. As world's expert, his beliefs about the risk function $b(\cdot)$ would not change upon learning the propensity score function $e(\cdot)$, as $e(\cdot)$ only reflected a nonexpert's beliefs. Hence the functions $b(\cdot)$ and $e(\cdot)$ are a priori independent. [Nonetheless, he would believe with high probability that the random variables $b(X)$ and $e(X)$ were positively correlated, knowing that the epidemiologist had read the expert literature on risk factors for MI.]

Robins and Ritov (1997) showed that once any Bayesian, cardiac expert or not, thoroughly queries the epidemiologist who selected $e(\cdot)$ about his reasoned opinions concerning $b(\cdot)$ (but not about $e(\cdot)$), the Bayesian will then have independent priors. The idea is that once you are satisfied that you have learned from the epidemiologist all he knows about $b(\cdot)$ that

you did not, you will have an updated prior for $b(\cdot)$. Your updated prior for $b(\cdot)$ cannot then change if you subsequently are told $e(\cdot)$. Hence, we could take as many Bayesians as you please and arrange it so all had $b(\cdot)$ and $e(\cdot)$ a priori independent. This last argument is quite general and applies to many settings.

Rejoinder

Olli Saarela,
David A. Stephens,
Erica E. M. Moodie,
and Marina B. Klein

The authors would like to thank all the discussants for their contributions and insights. In particular, the numerical results by Elliott and Little (from here on, EL) and Gustafson shed further light into the behavior of the different approaches for Bayesian causal inference. Mattei and Mealli (MM) point out a lack of clarity in our exposition for which we apologize and which we attempt to rectify below. Finally, Robins, Hernán and Wasserman (RHW) reiterate the incompatibility of a standard Bayesian formulation under independent prior specifications with the goal of uniformly consistent inference concerning the parameter of interest.

To recap: our paper attempts to demonstrate the formulation of Bayesian inference procedures for the now widely used marginal structural model (MSM) in particular, and for two-stage inverse weighting procedures in general. Due to space limitations in the original manuscript, we did not include a completely comprehensive summary of issues related to MSMs, presuming some familiarity on behalf of the reader. In particular, as MM comment, we did not discuss the MSM estimand at any length: we reiterate here that our estimand is merely the usual MSM estimand, that is, a parameter in a hypothesized marginal model relating the total effect of time-varying exposure to the counterfactual outcome (see Section 3.1). This parameter has a precise Bayesian interpretation through the de Finetti representation of a model for exchangeable observable quantities under specific assumptions concerning the data generating process – see Equation (6) in the main paper. However, crucially, these assumptions *are presumed not to hold for the actual data generating mechanism* at hand, requiring us to use reweighting methods. Our posterior predictive formulation of Bayesian inference follows for example, Walker (2010); to derive the estimator, we propose a utility function – specified as a loss function for a future predicted outcome – and compute its posterior (predictive) expectation. This computation is carried out using simulation based methods related to the weighted likelihood bootstrap of Newton and Raftery (1994), and also to the Bayesian bootstrap of Rubin (1981).

We now try to address in detail two specific points raised in the discussion, namely whether an inference procedure featuring IPT weights or propensity scores can ever be fully Bayesian (a question raised by both RHW and MM), and whether modeling of the treatment assignment mechanism is really preferable to modeling of the outcome and covariate processes (a question raised by MM, EL, and Gustafson).

1. Is it Bayesian?

First, we are in full agreement with RHW that from purely likelihood-based arguments, the treatment assignment mechanism plays no part in Bayesian inferences of marginal causal effects. Indeed, we noted that if the models are correctly specified, the parameters γ characterizing the treatment assignment mechanism are independent of the outcomes (Formula (4)). A dependency between these can arise if the outcome model is misspecified (Supplementary Appendix C), but in this case the balancing properties of the propensity scores would be lost, as has also been pointed out by McCandless et al. (2010) and Zigler et al. (2013). On the other hand, the marginal model parametrized in terms of θ is fully specified by the models for the outcome and covariate processes parametrized in terms of ϕ (Appendix), and inferences on θ can be obtained through marginalization if one is willing to model the full longitudinal data generating mechanism.

RHW proceed to make a convincing argument as to why one should not ignore propensity scores in favor of modeling the outcome and covariate processes; indeed this was also our motivation to study whether approximate posterior distributions for θ could be generated without specifying the models in terms of ϕ. RHW also point out – in line with the Robins–Ritov–Wasserman logic outlined in the Discussion – that no 'strict factorization based' estimator, such as that derived using Bayesian logic under prior independence assumptions, can perform uniformly adequately. As stated above, we concur with this view. However, through the use of a posterior predictive approach, we have constructed a fully Bayesian 'estimator' that is not strict factorization based.

We used a posterior predictive approach, where the assumed marginal model takes the role of a parametric utility function (cf. Walker, 2010). Our objective is to produce an estimator of the required marginal parameter by maximizing a posterior predictive expected utility; we feel that this is appropriately referred to as a Bayesian procedure. The unavailability of data suitable for computing the desired posterior predictive quantity requires us to adopt a simulation-based, Monte Carlo strategy for computing the necessary expectation integral. We have access to data collected in the

observational study, so may compute the corresponding observational posterior predictive distribution; we then use importance sampling ideas to rewrite the desired expected utility calculation in terms of this observationally derived quantity.

Although the change of measures or importance sampling argument has been previously used for deriving IPT weights by various authors – it is, of course, central to the construction of the classical Horvitz-Thompson estimator – to our knowledge it has not been used for this purpose in the Bayesian context, where it also suggests an inference procedure. Thus, we contend that the proposed approach is Bayesian in the same way that the weighted likelihood bootstrap of Newton and Raftery (1994) is; if we had actually observed a sample $v = (\tilde{x}, y, \tilde{z})$ under the hypothetical completely randomized setting $\mathcal{E}$, so that it would make sense to directly approximate the posterior predictive density $p(v_i^* \mid v, \mathcal{E})$ with the Bayesian bootstrap $\sum_{k=1}^{n} \pi_k \delta_{v_k}(v_i^*)$ where $\pi = (\pi_1, \ldots, \pi_n) \sim \text{Dirichlet}(1, \ldots, 1)$, we would obtain the maximum likelihood solution $\arg\max_\theta E[\ell(y_i^* \mid \tilde{z}_i^*, \theta) \mid v, \mathcal{E}] = \arg\max_\theta \sum_{i=1}^{n} \pi_i \ell(y_i \mid \tilde{z}_i, \theta)$. Here the weights π are a means to generate a probability distribution over the space of θ. Introducing the importance sampling weights to account for the non-random treatment assignments adds variability to the resulting estimator, depending on how strongly the treatment assignments depend on the covariates. Furthermore, the approach of Newton and Raftery (1994) is closely related to that of Rubin (1981), who also proposed Dirichlet sampling approximately to generate the posterior quantities of interest. Clearly, multinomial resampling (the standard nonparametric bootstrap) is a limiting special case of symmetric Dirichlet sampling (Rubin's Bayesian bootstrap) with prior parameters presumed taken to be infinitely large.

There are some drawbacks to this sampling approach to Bayesian inference; first, it does not allow direct specification of an informative prior for θ, although this can be added in afterwards by using the numerically obtained function over θ in place of a likelihood in Bayes' formula, as discussed in Section 3.4. Secondly, since the estimation procedure is formulated through an out-of-sample predictive criterion, conditioning on a given sample, it does not account for the variance reduction due to sample balance obtained through estimation of the IPT weights. This is because the weights are fixed to their best estimates based on the observed sample, rather than re-estimated in the predicted 'resamples', as in the frequentist bootstrap. Thus, while the proposed estimation procedure does not (and should not) add a variance component due to uncertainty in the estimation of the weights, it does not reduce the variance either, and thus results in a conservative variance estimator. (This was not apparent in our simulation setting of Section 5, but in other settings the variance reduction may be more substantial).

We can modify the weighting argument as follows to get closer to the correct frequentist – that is, repeated finite sample – properties. We note first that the expression $p_n(v_i^*) = \sum_{k=1}^{n} \pi_k \delta_{v_k}(v_i^*)$ for the posterior predictive density $p(v_i^* \mid v, \mathcal{O})$ implies also an expression for any of the conditional distributions

$$p_n(z_{ij}^* \mid \tilde{z}_{i(j-1)}^*, \tilde{x}_{ij}^*) = \frac{\sum_{k=1}^{n} \pi_k \delta_{(\tilde{z}_{kj}, \tilde{x}_{kj})}(\tilde{z}_{ij}^*, \tilde{x}_{ij}^*)}{\sum_{k=1}^{n} \pi_k \delta_{(\tilde{z}_{k(j-1)}, \tilde{x}_{kj})}(\tilde{z}_{i(j-1)}^*, \tilde{x}_{ij}^*)},$$

$j = 1, \ldots, m$, which are thus fully determined by π and v. In practice it is not feasible to directly use the non-parametric expression for estimating the treatment assignment probabilities, and a parametric specification $p(z_{ij}^* \mid \tilde{z}_{i(j-1)}^*, \tilde{x}_{ij}^*, \gamma_j, \mathcal{O})$ would be used instead. However, the parameters γ_j themselves may in turn be estimated using the weighted likelihood bootstrap, as

$$\begin{aligned}
&\arg\max_{\gamma_j} E\left[\log p(z_{ij}^* \mid \tilde{z}_{i(j-1)}^*, \tilde{x}_{ij}^*, \gamma_j, \mathcal{O}) \mid v, \mathcal{O}\right] \\
&\quad = \arg\max_{\gamma_j} \sum_{i=1}^{n} \pi_i \log p(z_{ij} \mid \tilde{z}_{i(j-1)}, \tilde{x}_{ij}, \gamma_j, \mathcal{O}) \\
&\quad \equiv \widehat{\gamma}_j(v; \pi).
\end{aligned}$$

Since $\widehat{\gamma}_j(v; \pi)$ is taken to be a sampled value from the posterior distribution $p(\gamma_j \mid v)$, this motivates approximating $p_n(z_{ij}^* \mid \tilde{z}_{i(j-1)}^*, \tilde{x}_{ij}^*)$ parametrically with $p(z_{ij}^* \mid \tilde{z}_{i(j-1)}^*, \tilde{x}_{ij}^*, \widehat{\gamma}_j(v; \pi), \mathcal{O})$. A similar argument would apply for the parameters α_j, $j = 1, \ldots, m$, specifying the marginal treatment assignment probabilities.

Considering now the original utility function given a realization of π, we get

$$\begin{aligned}
&E[\ell(y_i^* \mid \tilde{z}_i^*, \theta) \mid v, \mathcal{E}] \\
&\quad = \int_{v_i^*} \ell(y_i^* \mid \tilde{z}_i^*, \theta) \frac{\prod_{j=1}^{m} p_n(z_{ij}^* \mid \tilde{z}_{i(j-1)}^*)}{\prod_{j=1}^{m} p_n(z_{ij}^* \mid \tilde{z}_{i(j-1)}^*, \tilde{x}_{ij}^*)} p_n(v_i^*) \, \mathrm{d}v_i^* \\
&\quad = \sum_{i=1}^{n} w_i \pi_i \ell(y_i \mid \tilde{z}_i, \theta),
\end{aligned}$$

where the non-parametrically specified weights would in practice be replaced with the parametric versions

$$w_i = \frac{\prod_{j=1}^{m} p(z_{ij} \mid \tilde{z}_{i(j-1)}, \widehat{\alpha}_j(v; \pi), \mathcal{O})}{\prod_{j=1}^{m} p(z_{ij} \mid \tilde{z}_{i(j-1)}, \tilde{x}_{ij}, \widehat{\gamma}_j(v; \pi), \mathcal{O})}.$$

The modified computational algorithm now involves re-estimating the IPT weights at each realization of π. Using the multinomial distribution instead of the Dirichlet distribution would (with flat priors) reproduce the frequentist bootstrap. Simulation results for the modified Bayes/Dirichlet estimator are reported in Rejoinder Table 1; the same simulation setting as in Section 5.2 was employed. These results resemble those for the frequentist IPT weighted estimator combined with the adjusted sandwich variances (Table 1), with some undercoverage visible when large weights are present. Since the frequentist bootstrap, which also involves re-estimation of the weights in each resample, does not show such undercoverage, we take this to be related to the properties of the Dirichlet and multinomial distributions. For comparison, in the real data example of Section 6.2, the point estimate and its standard error with the re-estimated weights were 0.302 and 0.339, respectively, also showing slightly smaller variability compared to the previous results in Table 3. Further simulations would be needed to investigate the properties of the modified estimation procedure under different settings.

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
$b = 0; \theta_2 = -0.247$	Bayes/Dirichlet	-0.256	-3.613	0.108	0.105	-3.403	94.4
$b = 0.15; \theta_2 = -0.569$	Bayes/Dirichlet	-0.577	-1.405	0.143	0.133	-7.052	93.3
$b = 0.3; \theta_2 = -0.777$	Bayes/Dirichlet	-0.762	1.915	0.212	0.175	-17.635	90.2

2. Is Modeling of the Treatment Assignment Mechanism Preferable to Modeling of the Outcome and Covariate Processes?

MM point out that also IPT weighted approaches to inference require plenty of identifying and modeling assumptions; in particular, we need to assume that the models for the treatment assignment mechanism are correctly specified, all the relevant confounders are measured, and that the positivity/overlap condition is not violated. Whilst noting that these assumptions are arguably less restrictive than the requirement to construct a completely correctly specified longitudinal model for exposure and outcome, we fully agree that the practical evaluation of these properties is crucial, and this is equally the case for the inference procedure proposed herein. However, because the practical issues are the same, and have been discussed extensively and in detail by other authors (e.g., Cole and Hernan, 2008; Xiao, Moodie, Abrahamowicz, 2013, and the references given by MM), we did not discuss them at length. It is certainly true that especially in longitudinal settings the variability of the estimated weights can become excessive, and lead to substantial loss of precision compared to likelihood-based inferences based on modeling the covariate and outcome processes, as is also demonstrated in the simulation study by EL. However, in terms of bias, as noted in the previous section, a modeling strategy based on finite-dimensional parametrizations is arguably likely to be more successful for the treatment assignment mechanism than for the distribution of all longitudinal covariates (i.e. the confounders and mediators), in particular if the longitudinal covariates are high-dimensional.

Other issues were also raised; for instance, many of the discussants preferred the more conventional Bayesian formulation where potential outcomes are considered as missing data. We did not rely on potential outcomes notation to formulate the causal estimands and estimators, but rather opted to define these in terms of the observational and experimental measures for exchangeable observable sequences via de Finetti's representation. We do not wish to enter into a debate on the respective merits of different notational systems for essentially equivalent approaches to causal modeling; the causal inference problem may be formulated alternatively as a missing data problem, or a prediction problem (cf. Greenland, 2012).

References

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

Greenland, S. (2012). Causal inference as a prediction problem: assumptions, identification and evidence synthesis. In Berzuini, C., Dawid, A. P., and Bernardinelli, L. (eds), *Causality: Statistical Perspectives and Applications*, pp. 43–58. New York: Wiley.

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.

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.

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

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

Xiao, Y., Moodie, E. E. M., and Abrahamowicz, M. (2013). Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods* **2**, 1–20.

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.