Published in final edited form as:

Am Stat. 2016 March 31; 70(1): 47–54. doi:10.1080/00031305.2015.1111260.

The Central Role of Bayes' Theorem for Joint Estimation of Causal Effects and Propensity Scores

Corwin Matthew Zigler

Department of Biostatistics, Harvard T.H. Chan School of Public Health

Abstract

Although propensity scores have been central to the estimation of causal effects for over 30 years, only recently has the statistical literature begun to consider in detail methods for Bayesian estimation of propensity scores and causal effects. Underlying this recent body of literature on Bayesian propensity score estimation is an implicit discordance between the goal of the propensity score and the use of Bayes theorem. The propensity score condenses multivariate covariate information into a scalar to allow estimation of causal effects without specifying a model for how each covariate relates to the outcome. Avoiding specification of a detailed model for the outcome response surface is valuable for robust estimation of causal effects, but this strategy is at odds with the use of Bayes theorem, which presupposes a full probability model for the observed data that adheres to the likelihood principle. The goal of this paper is to explicate this fundamental feature of Bayesian estimation of causal effects with propensity scores in order to provide context for the existing literature and for future work on this important topic.

Keywords

Bayesian estimation; Causal inference; Model feedback

1 INTRODUCTION

The propensity score (Rosenbaum and Rubin, 1983) plays a central role in a variety of methods for estimating causal effects with observational data. While the usefulness of propensity scores for Bayesian inference was motivated decades ago (Rubin, 1985), the promise of modern Bayesian analysis tools has spawned renewed interest in Bayesian estimation of causal effects with propensity scores (Hoshino, 2008; McCandless et al., 2009; An, 2010; McCandless et al., 2010; Kaplan and Chen, 2012; Zigler et al., 2013; McCandless et al., 2012; Zigler and Dominici, 2014; Saarela et al., 2015).

The literature on propensity scores is vast (see useful reviews in DAgostino (1998); Lunceford and Davidian (2004); Rubin (2008); Stuart (2010)), but a commonality among virtually all propensity score methods is that they unfold in two stages: the propensity score is estimated as the conditional probability of receiving treatment (conditional on covariates)

and then outcomes are compared (via matching, weighting, or subclassification) among observations with similar values of the estimated propensity score. These two stages are sometimes referred to as approximating the “design” and “analysis” of a hypothetical randomized study (Ho et al., 2007; Rubin, 2008; Rubin and Waterman, 2006). A successful design stage ensures that outcome comparisons in the analysis stage are conducted among observations that are comparable with respect to observed factors that would otherwise confound estimates of the treatment effect.

Traditional estimation of causal effects with propensity scores follows this two-stage framework: outcome comparisons in the analysis stage are made conditional on point estimates of the propensity score obtained from the design stage. In light of this inherent two-stage nature, the term “Bayesian propensity score analysis” is not entirely clear. Bayesian inference has been considered in the design stage, with Rosenbaum and Rubin (1983) noting that propensity score estimates can be regarded as posterior-predictive probabilities of treatment assignment. However, Rosenbaum and Rubin (1983) did not consider how to incorporate the posterior-predictive distribution of propensity scores into the analysis stage. Rubin (1985) describes Bayesian inference for the analysis stage that conditions entirely on estimates of the propensity score, without regard to how the propensity scores were estimated in the design stage. In essence, the use of Bayesian methods in either the design stage or the analysis stage is straightforward, but debate remains about how to combine these two stages into a single Bayesian analysis. In McCandless et al. (2009) a “Bayesian propensity score analysis” is proposed to use the mechanics of Bayesian estimation to unite the design and analysis stages into a single procedure, but Zigler et al. (2013) showed that “feedback” of information between the two stages could render the procedure unable to reliably estimate causal effects. More recently, Saarela et al. (2015) proposed a Bayesian approach that uses propensity scores to estimate the parameters of a marginal structural model, which incited interesting discussions about whether the approach could be truly “Bayesian” (Robins et al., 2015). A variety of other methods have been proposed and described with terms such as “quasi-Bayesian” or “two-step Bayesian” to employ the Bayesian inferential paradigm in propensity score and causal effect estimation while maintaining the conceptual separation of design and analysis stages (McCandless et al., 2010; Kaplan and Chen, 2012; Zigler and Dominici, 2014). The purpose of this article is to clarify, against the backdrop of existing literature on “Bayesian propensity score analysis,” the fundamental issues at play when attempting to conduct Bayesian inference amid the two-stage nature of estimating causal effects with propensity scores.

A key element of the research in this area pertains to uncertainty quantification. Since traditional non-Bayesian estimation procedures treat estimated propensity scores as fixed in the analysis stage, much existing literature on Bayesian methods has focused on the ability to propagate uncertainty in estimation of propensity scores into estimates of causal effects (McCandless et al., 2009; An, 2010; Kaplan and Chen, 2012). In this regard, the existing work on “Bayesian propensity scores” contributes to a diverse literature on variance estimation with propensity scores, which includes, among other results, the feature that conditioning on the estimated propensity score is more efficient than knowing the true propensity score (Rubin and Thomas, 1996; Wooldridge, 2002; Brumback, 2009). The

interest here is broader. Rather than adopt a narrow focus on one feature such as the efficiency tradeoffs of acknowledging propensity score uncertainty, the goal is to clarify the fundamental issues that underly existing attempts to incorporate propensity scores into broader Bayesian methodology. Improved understanding of the fundamental interplay between propensity scores and Bayesian inference is an avenue to unite the established virtues of propensity scores with the benefits of modern Bayesian methodology. Such a unification can motivate new methods targeting new research questions, with important progress recently appearing for using propensity scores in Bayesian estimation of parameters in latent variable models (Hoshino, 2008), in Bayesian models that use propensity scores for combining data from multiple sources (McCandless et al., 2012), and in methods for Bayesian model averaging to address uncertainty in the propensity score model specification (Zigler and Dominici, 2014). Improved understanding of the complexities of Bayesian inference with propensity scores will permit continued progress towards these and other methodological goals.

Despite important progress, there exists a fundamental difficulty with unifying the two-stage nature of estimating causal effects with propensity scores into a coherent Bayesian analysis. This difficulty and its implications for estimating causal effects has been alluded to in existing literature in various ways, both conceptually (e.g., as a separation between design and analysis stages) and technically (e.g., as arguments about uniform consistency of Bayes estimators that do or do not assume independent prior knowledge of the propensity score) (Kaplan and Chen, 2012; Robins et al., 1997, 2015). The goal of this paper is not to propose new methodology nor advocate for the superiority of Bayesian approaches, but rather to explicate the fundamental features of Bayesian estimation of causal effects with propensity scores in a manner accessible to the statistician who is familiar with propensity score estimation and with the basic tenets of Bayesian inference. The hope is to provide context for the existing literature and for future work on this important topic.

2 ESTIMATION OF CAUSAL EFFECTS WITH PROPENSITY SCORES

Consider n observational units that have been randomly sampled from a population of interest. For each unit, consider a continuous, categorical, ordinal, or time-to-event outcome, Y_i , for $i = 1, 2, \dots, n$. Let X_i be a binary variable with $X_i = 1$ denoting exposure to a treatment of interest and $X_i = 0$ otherwise. The goal is to estimate the causal effect of exposure to $X = 1$ (vs. $X = 0$) on Y , which is typically complicated in observational settings because assignment to treatment is not randomized. Let $C_i = C_{i1}; C_{i2}, \dots, C_{ip}$ denote a vector of p pre-treatment variables that simultaneously relate to the receipt of treatment and to the outcome of interest, that is, let C_i be the vector of confounders measured for $i = 1, 2, \dots, n$.

Causal inference literature, including that involving propensity scores, is most formally rooted in the notion of *potential outcomes* representing, for each i , the outcome that would potentially be observed if unit i were assigned either $X_i = 1$ or $X_i = 0$ (Rubin, 1978). Definition of potential outcomes is one integral feature of the so-called “Rubin Causal Model” (RCM). While all methods considered here are founded in the RCM, potential outcomes notation is avoided for ease of exposition, with references to potential outcomes only when necessary for clarification.

The estimand of interest is the average causal effect, denoted with Δ and defined as:

$$\Delta \equiv E_C[E[Y_i|X_i=1, C_i] - E[Y_i|X_i=0, C_i]]. \quad (1)$$

Expression (1) is interpretable as a causal effect of X on Y under the assumption of *strongly ignorable treatment assignment* that, conditional on C_i , assignment to X_i is conditionally independent of potential outcomes that would be observed under either possible value of X_i (Rosenbaum and Rubin, 1983). The remainder of this paper assumes strongly ignorable treatment assignment, that is, that there is no unmeasured confounding. Note that this assumption cannot be verified empirically, and its justification relates to confidence that C_i captures all determinants of assignment to $X = 0$ vs. $X = 1$ that have bearing on Y .

2.1 The rationale for using propensity scores to estimate causal effects

Many statistical methods are available to estimate causal effects such as (1) from observational data. Most commonly, inference follows from estimation of parameters in a model for the outcome response surface that relates observed Y to observed values of X and C , for example, a linear regression model. Propensity score methods typically do not rely on parametric models for the response surface for Y . Instead, the multivariate vector C_i is condensed into a scalar-valued confounder summary, namely, the propensity score, that can be used to generate unconfounded treatment comparisons. The propensity score for the i^{th} unit is the conditional probability of exposure to $X_i = 1$, conditional on C_i : $e_i \equiv p(X_i = 1|C_i)$. In observational studies, the propensity score is unknown and must be estimated. Denote the estimated propensity score with $\hat{e}_i$.

The rationale for estimating causal effects with propensity scores rests on the result that e_i is a *balancing score* so that $X_i \perp C_i|e_i$; conditional on the propensity score, the treatment is independent of the covariates. Effectively balancing the covariates in this manner corresponds to the successful “design” of an experiment where treated and untreated observations with similar values of e_i have similar covariate profiles. This property, coupled with the ignorability assumption, allows average outcome comparisons among observations with the same value of e_i to serve as estimates of the average causal effect at that value of e_i . Thus, the causal effect Δ in (1) can be rewritten as:

$$\Delta \equiv E_e[E[Y_i|X_i=1, e_i] - E[Y_i|X_i=0, e_i]]. \quad (2)$$

There are a variety of methods available for estimating (2) with propensity scores whereby the estimated quantities $\hat{e}_i$ are obtained and then causal effects are estimated by treating $\hat{e}_i$ as a covariate, matching or subclassifying units by $\hat{e}_i$, or weighting by a function of $\hat{e}_i$ to calculate $E_{\hat{e}}[E[Y_i|X_i=1, \hat{e}_i] - E[Y_i|X_i=0, \hat{e}_i]]$ (Rosenbaum and Rubin, 1983; Robins et al., 2000; Lunceford and Davidian, 2004; Stuart, 2010).

Motivation for using propensity score methods are many. Relying on $\hat{e}_i$ allows researchers to avoid reliance on a parametric model specification for how Y relates to each element of a (possibly high dimensional) covariate vector, C . That is, adjustment for the propensity score

is not meant to reflect a model for the data Y ; it is simply meant to facilitate comparisons between $[Y/X = 1]$ and $[Y/X = 0]$ that are not confounded by C so that such contrasts can estimate causal effects. This can be particularly beneficial in settings where more is known about the mechanism governing assignment to X than about the relationships between C , X , and Y , or in settings where the treatment is common and the outcome is rare (Braitman and Rosenbaum, 2002; Cepeda et al., 2003). Estimating $\hat{e}_i$ in the design stage also has the benefit of permitting the analyst to check whether the design of the hypothetical trial is successful. Analysts can check for gross imbalances in C among treated and untreated units with similar values of $\hat{e}_i$, which would indicate that confounding has not been completely eliminated. Furthermore, estimates of $\hat{e}_i$ can be used to gauge propensity score “overlap,” that is, the existence of treated and untreated units that have comparable covariate profiles. Units identified to have $\hat{e}_i$ that substantially differs from every observation in the opposite treatment group are not appropriate for making causal inference without some form of extrapolation (Ho et al., 2007; Crump et al., 2009). To focus presentation, detailed consideration of these elements of propensity score analysis is not pursued, but useful discussion appears in Rubin (2008); DAGostino (1998); Rubin and Waterman (2006).

2.2 Propensity Scores vs. Models for Potential Outcomes

Propensity score methods are used to facilitate average observed-data comparisons between one set of units (those with $X = 1$) and another set of units (those with $X = 0$) in a manner that is not confounded. As an alternative to propensity score methods, the underlying theory of potential outcomes has motivated a vast and well-established literature on Bayesian causal inference that does not rely on propensity scores, but rather employs the Bayesian inferential paradigm with models specified for the joint distribution of potential outcomes (e.g., Imbens and Rubin (1997); Jin and Rubin (2008); Schwartz et al. (2011)). Given the fundamental problem of causal inference that potential outcomes under treatments not received are missing, this literature makes use of the mechanics of Bayesian missing data imputation to estimate potential outcomes under treatments not received. This typically involves predicting missing potential outcomes for a given unit conditional on that unit’s observed potential outcome, which requires, for example, the need to consider the non-identified correlation between potential outcomes under opposite treatments. While modeling potential outcomes can be regarded as a refinement over strategies that model only observed outcomes, these approaches still rely on parametric model specifications for (potential) outcomes, and thus do not enjoy the purported benefits of propensity-score methods described in Section 2.1. The role of Bayesian inference when using such models for potential outcomes is well understood. In contrast, the precise role of Bayesian inference procedures in the context of methods that use propensity scores without specifying a model for (potential) outcomes warrants clarification. This manuscript compares Bayesian estimation approaches with traditional non-Bayesian approaches when using the exact same model formulations relying on propensity scores. The consideration of Bayesian approaches simply pertains to whether the parameters of these models (which remain the same regardless of estimation strategy) can be estimated reliably.

2.3 Models for estimating propensity scores and causal effects

This paper considers approaches that rely on 1) a “propensity score model” for the treatment assignment mechanism and 2) and a propensity score “outcome model” that models Y as a function of the propensity score, e_i . Specifying the outcome model to depend on the propensity score is the key point of contrast with a procedure that would instead model Y as a function of the covariates, C_i . While the majority of the following discussion focuses on the specification of both the propensity score and outcome models, note that an outcome model is not strictly required for estimating with propensity scores. Non-model-based strategies are discussed in Section 5. Throughout, let θ represent parameters indexing models for observed data, with different subscripts indicating that parameter vectors are distinct depending on different model specifications.

Let $m_X(X_i, C_i; \theta_{X/C})$ denote the propensity score model, that is, a model for $p(X_i = 1/C_i)$, indexed by a parameter $\theta_{X/C}$. Estimates from this model yield $\hat{e}_i$. Note that specification of $m_X(X_i, C_i; \theta_{X/C})$ can vary in flexibility from commonly-used logistic regression to modern machine-learning specifications such as boosted regression trees (McCaffrey et al., 2004). Henceforth, consider augmented notation to explicate the link between the propensity score and $m_X(X_i, C_i; \theta_{X/C})$ and denote the propensity score and its estimate with $e_i(C_i; \theta_{X/C})$ and $\hat{e}_i(C_i; \hat{\theta}_{X/C})$, respectively. When $m_X(X_i, C_i; \theta_{X/C})$ accurately reflects the treatment assignment mechanism, $\hat{e}_i(C_i; \hat{\theta}_{X/C})$ enjoys the property of a balancing score.

Denote the outcome model with $m_Y(Y_i, X_i, e_i(C_i; \theta_{X/C}); \theta_{Y/X,e})$, which could simply specify a regression model with covariate adjustment for $e_i(C_i; \theta_{X/C})$, or correspond to a flexible parameterization that estimates a different causal effect for different subgroups (e.g., quintiles or matched subclasses) of observations having similar values of $e_i(C_i; \theta_{X/C})$.

Traditional estimation of would specify $m_X(X_i, C_i; \theta_{X/C})$ and estimate $\theta_{X/C}$ to obtain $\hat{e}_i(C_i; \hat{\theta}_{X/C})$ and then use $m_Y(Y_i, X_i, \hat{e}_i(C_i; \hat{\theta}_{X/C}); \theta_{Y/X,e})$ to estimate $\theta_{Y/X,e}$ and obtain an estimate of . That is, estimation of causal effects with the outcome model is conducted conditional on the estimated propensity score.

3 Bayesian Estimation of Causal Effects with Propensity Scores

3.1 Combining the Propensity Score and Outcome Models into a Single “Likelihood”

For Bayesian inference, a model, however flexible, must be specified for the joint probability distribution of the data and parameters, here $(\mathbf{X}, \mathbf{Y}, \mathbf{C}, \theta)$, where boldface letters represent vectors and matrices for observed quantities for all n units and θ denotes a fixed but unknown generic parameter vector. As noted in Section 2.2, an alternative approach would be to specify a joint distribution for potential outcomes directly, in which case issues analogous to those described below would persist. For application of Bayes theorem, this joint distribution is factored into an expression for the likelihood, $p(\mathbf{X}, \mathbf{Y}, \mathbf{C}|\theta)$, describing how the data are generated conditional on the true θ , and a prior distribution, $p(\theta)$, describing *a priori* beliefs about possible values of the true θ . Information obtained in the data updates $p(\theta)$ to yield a posterior distribution of $p(\theta|\mathbf{X}, \mathbf{Y}, \mathbf{C})$ that can be used to estimate . The posterior distribution of can be expressed as:

$$p(\Delta|\mathbf{X}, \mathbf{C}, \mathbf{Y}) \propto \int \Delta(\mathbf{X}, \mathbf{Y}, \mathbf{C}, \theta) \prod_{i=1}^n p(X_i, Y_i, C_i|\theta) p(\theta) d\theta, \quad (3)$$

where $(\mathbf{X}, \mathbf{Y}, \mathbf{C}, \theta)$ denotes that, conditional on θ , the quantity is a deterministic function of $(\mathbf{X}, \mathbf{Y}, \mathbf{C}, \theta)$, and $\prod p(X_i, Y_i, C_i|\theta) p(\theta)$ is proportional to the posterior distribution $p(\theta|\mathbf{X}, \mathbf{Y}, \mathbf{C})$.

The key modeling steps for Bayesian analysis are specification of a prior distribution, $p(\theta)$, and a model for $p(X_i, Y_i, C_i|\theta)$ constituting the likelihood function. Throughout, take the marginal distribution of $\mathbf{C}$ as its empirical distribution, and omit this distribution from likelihood expressions to simplify notation. The factorization $p(X_i, Y_i, C_i|\theta) = p(X_i|C_i, \theta_{X|C}) p(Y_i|X_i, C_i, \theta_{Y|X,C})$ seems to correspond naturally to the two stages of estimating causal effects with propensity scores; $p(X_i|C_i, \theta_{X|C})$ is the propensity score model and $p(Y_i|X_i, C_i, \theta_{Y|X,C})$ is the outcome model. Should a researcher be confident in specifying the outcome model, $p(Y_i|X_i, C_i, \theta_{Y|X,C})$, Bayesian inference for could follow from standard Bayesian estimation of the posterior distribution of $\theta_{Y|X,C}$. Assuming *a priori* independence between $\theta_{Y|X,C}$ and $\theta_{X|C}$, the propensity score model would play no role due to the likelihood principle that any likelihood function proportional to $p(Y_i|X_i, C_i, \theta_{Y|X,C})$ would yield equivalent inference for $\theta_{Y|X,C}$ and, in turn (Robins et al., 1997, 2015).

However, propensity score methods do not rely on a model for $p(Y_i|X_i, C_i, \theta_{Y|X,C})$, in fact, they are specifically motivated by a desire to *avoid* specification of such a model. Employing the propensity score to estimate would imply an outcome model as a function of $e_i(C_i, \theta_{X|C})$, which would correspond to the following likelihood expression:

$$\prod_{i=1}^n p(X_i|C_i, \theta_{X|C}) p(Y_i|X_i, e_i(C_i, \theta_{X|C}), \theta_{Y|X,e}), \quad (4)$$

where $p(X_i|C_i, \theta_{X|C})$ and $p(Y_i|X_i, e_i(C_i, \theta_{X|C}), \theta_{Y|X,e})$ are modeled with $m_X(X_i, C_i, \theta_{X|C})$ and $m_Y(Y_i, X_i, e_i(C_i, \theta_{X|C}), \theta_{Y|X,e})$, respectively. Note that the parameter $\theta_{Y|X,e}$ is distinct from $\theta_{Y|X,C}$.

Bayesian estimation of causal effects with (4) was proposed in McCandless et al. (2009), but shown in Zigler et al. (2013) to yield biased estimates of causal effects without further adjustments. That is, the posterior distribution in (3) cannot, in general, be recovered with a likelihood specified as in (4).

3.2 The fundamental problem of joint Bayesian estimation of propensity scores and causal effects

Because the propensity score is not meant to recover $p(Y_i|X_i, C_i, \theta_{Y|X,C})$, the fundamental problem for using Bayes theorem to combine estimation of propensity scores with estimation of causal effects can be succinctly stated in the following expression:

$$p(\mathbf{X}, \mathbf{Y}, \mathbf{C}|\theta) = \prod_{i=1}^n p(X_i, Y_i, C_i|\theta) \neq \prod p(X_i|C_i, \theta_{X|C})p(Y_i|X_i, e_i(C_i, \theta_{X|C}), \theta_{Y|X,e}), \quad (5)$$

i.e., a model for how $e_i(C_i, \theta_{X|C})$ relates to Y_i , however flexible, cannot in general serve the same purpose in Bayesian analysis as a model for the joint distribution $p(\mathbf{X}, \mathbf{Y}, \mathbf{C}|\theta)$. Strictly speaking, the rightmost expression in (5) does not even correspond to a likelihood function at all, as it is specifically designed to avoid specification of key features of the joint distribution of the data, namely, the covariate-outcome response surface. Thus, use of the expression in (4) does not correspond to a valid use of Bayes theorem because it does not correspond to a coherent factorization of the joint distribution of the observed data, conditional on θ .

One practical consequence of this fundamental issue was described in Zigler et al. (2013) as “feedback” of information from $p(Y_i|X_i, e_i(C_i, \theta_{X|C}), \theta_{Y|X,e})$ into estimation of $\theta_{X|C}$ and, in turn, $e_i(C_i, \theta_{X|C})$. This phenomenon was shown to provide invalid inferences for $\theta_{X|C}$, which distorts the balancing-score property of $e_i(C_i, \theta_{X|C})$ and yields incorrect estimates of $\theta_{X|C}$, confirming that an expression such as that in (4) cannot be used to recover the posterior distribution of $\theta_{X|C}$ in (3). In short, failure of expression (4) to adhere to a coherent probability model leads to incorrect estimates of the propensity score, improper confounding adjustment, and inconsistent estimates of causal effects. However, Zigler et al. (2013) also show that augmenting the outcome model to be parameterized in terms of both $e_i(C_i, \theta_{X|C})$ and C_i directly does permit estimation of $\theta_{X|C}$ with (3). This strategy implies a joint distribution for $(\mathbf{Y}, \mathbf{X}, \mathbf{C})$, and also benefits from accurate characterization of $e_i(C_i, \theta_{X|C})$ for estimating causal effects. Augmenting the outcome model to include both $e_i(C_i, \theta_{X|C})$ and individual covariate adjustment is advocated for a variety of reasons (Stuart, 2010), and has been employed in Bayesian propensity score methods in Zigler and Dominici (2014).

3.3 Two-Step Bayesian Procedures

Various methods described as “quasi-Bayesian,” “approximately-Bayesian,” or “two-step Bayesian” have been recently proposed to combine estimation of the propensity score with that of causal effects without appealing to Bayes theorem to unify these two stages with a joint likelihood (Hoshino, 2008; McCandless et al., 2010; Kaplan and Chen, 2012; Zigler and Dominici, 2014). These methods have been described as “cutting the feedback” between the propensity score and outcome stages (Lunn et al., 2009), and represent a special case of so-called “modularization” in Bayesian inference (Liu et al., 2009). These strategies use the posterior distribution $p(\theta_{X|C}|\mathbf{X}, \mathbf{C})$ to obtain posterior-predictive simulations of the propensity scores, without conditioning on Y or other quantities in the outcome model.

Denote posterior-predictive propensity score simulations with $e_i(C_i, \theta_{X|C}^*)$. Using simulations of $e_i(C_i, \theta_{X|C}^*)$ to estimate causal effects can be denoted with:

$$\begin{aligned}
 p(\Delta|\mathbf{X}, \mathbf{C}, \mathbf{Y}) \propto & \\
 \int \Delta(\mathbf{X}, \mathbf{Y}, \mathbf{e}(\mathbf{C}, \theta_{\mathbf{X}|\mathbf{C}}^*), \theta_{\mathbf{Y}|\mathbf{X}, \mathbf{e}^*}) p(\mathbf{Y}|\mathbf{X}, \mathbf{e}(\mathbf{C}, \theta_{\mathbf{X}|\mathbf{C}}^*), \theta_{\mathbf{Y}|\mathbf{X}, \mathbf{e}^*}) & \\
 \times \left[\int \mathbf{e}(\mathbf{C}, \theta_{\mathbf{X}|\mathbf{C}}) p(\theta_{\mathbf{X}|\mathbf{C}}|\mathbf{X}, \mathbf{C}) d\theta_{\mathbf{X}|\mathbf{C}} \right] & \\
 \times p(\theta_{\mathbf{Y}|\mathbf{X}, \mathbf{e}^*}) d\theta_{\mathbf{Y}|\mathbf{X}, \mathbf{e}^*}, & \quad (6)
 \end{aligned}$$

where $\theta_{\mathbf{Y}|\mathbf{X}, \mathbf{e}^*}$ is distinct from $\theta_{\mathbf{Y}|\mathbf{X}}$ in expression (4) because $\theta_{\mathbf{Y}|\mathbf{X}, \mathbf{e}^*}$ indexes a model that is conditional on the posterior-predictive propensity score based only on $p(\theta_{\mathbf{X}|\mathbf{C}}|\mathbf{X}, \mathbf{C})$, without conditioning on $\mathbf{Y}$.

Note that the interior integral in expression (6) with respect to $\theta_{\mathbf{X}|\mathbf{C}}$ represents the posterior-predictive propensity score distribution, which is obtained independently of any model specification for $\mathbf{Y}$. Thus, estimating causal effects with (6) maintains the two-stage nature of estimating causal effects with propensity scores and does not appeal to Bayes theorem to combine these stages, nor does it attempt to model the entire joint distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{C})$.

This procedure parallels that of multiple imputation, where, say m samples of $e_i(C_i, \theta_{\mathbf{X}|\mathbf{C}}^*)$ are obtained according to the posterior predictive distribution in the interior integral of expression (6), treated as a “distributional constant,” and iteratively plugged in to the outcome model to obtain m samples from a posterior distribution of while reflecting uncertainty in the estimated propensity score. Another way to frame the expression in (6) is to regard $\theta_{\mathbf{X}|\mathbf{C}}^*$ in the outcome model as an unknown parameter with an informative prior distribution, where that prior distribution is precisely the posterior distribution $p(\theta_{\mathbf{X}|\mathbf{C}}|\mathbf{X}, \mathbf{C})$ obtained during the design stage of the analysis. This would imply a prior for the outcome model that is dependent on the parameters of the propensity score model. It also bears noting that these methods strictly adhere to one philosophical motivation that the propensity score, as a tool meant to mimic the design of a randomized study, should be estimated separately from the analysis of causal effects (Rubin, 2008).

4 ILLUSTRATION WITH SIMULATED DATA

To illustrate the practical consequences of the issues raised in Section 3 in the context of one common propensity score implementation, a simple simulation study was conducted. The goal is to compare the Bayesian and two-step Bayesian procedures outlined in Section 3 to traditional propensity score estimation, not to investigate whether any particular method proves superior, but rather to benchmark against the standard estimation approach that has been evaluated in the literature over the past several decades. The simulation study considers these approaches when a model is used to estimate the casual effect in strata of the population that are sub classified according to quintiles of the estimated propensity score. For 1000 Monte Carlo replicates, data sets are simulated with binary X_i , Y_i and $p = 6$ covariates $C_i = C_{i1}, C_{i2}, \dots, C_{i6}$ for $i = 1, 2, \dots, 1000$. For each replicated data set, X_i and Y_i are simulated from Bernoulli distributions with probabilities of success generated from logistic regressions as follows:

$$\text{logit}(p(X_i=1|C_i))=0.1C_{i1}+0.2C_{i2}+0.3C_{i3}+0.4C_{i4}+0.5C_{i5}+0.6C_{i6} \quad (7)$$

$$\text{logit}(p(Y_i=1|X_i, C_i))=0.6C_{i1}+0.5\exp\{C_{i2}-1\}+0.4C_{i3}+0.3\exp\{C_{i4}-1\}+0.2|C_{i5}|+0.1|C_{i6}|$$

(8)

Following the approach described in Section 2.3, the simulated data sets are analyzed with $m_X(X_i, C_i; \theta_{X/C})$ specified as a logistic regression with a main effect for each component of C_i , which is shown in Table 1. Thus, the propensity score model used for analysis is correctly specified in accordance with the true treatment assignment mechanism in (7) (except that the analysis model permits a nonzero intercept). This is done to ensure that any evident bias in causal effect estimation with the strategies below is not due to the model specification for the propensity score. Data are analyzed with two different outcome models. Both models are logistic regressions that include a main effect for X_i and dummy variables indicating membership in quintiles of $e_A(C_i; \theta_{X/C})$. Outcome Model #1 adjusts only for $e_A(C_i; \theta_{X/C})$ quintile membership, while Outcome Model #2 augments this adjustment for $e_A(C_i; \theta_{X/C})$ quintile membership with linear adjustment for each individual C_i . Both outcome models are shown in Table 1. Note that the linear adjustment term in Outcome Model #2 does not reflect the transformed C_i used for data generation in (8).

Each simulated data set is analyzed with five strategies: Strategy A1 uses a traditional sequential approach described in Section 2.3 where maximum likelihood estimates are used to estimate $\hat{e}_A(C_i; \theta_{X/C})$ and Outcome Model #1 is used to obtain a maximum likelihood estimate of τ , conditional on $\hat{e}_A(C_i; \theta_{X/C})$; Strategy A2 uses the same traditional sequential approach but with Outcome Model #2; Strategy B1 uses a Bayesian approach with the likelihood in (4) and Outcome Model #1, which does not correspond to a valid use of Bayes theorem; Strategy B2 uses the Bayesian approach described in Section 3.2 that uses Outcome Model #2; Strategy C uses a two-step Bayesian approach described in Section 3.3 with Outcome Model #1.

Results from the simulated data sets are summarized in Figure 1, which depicts the mean and (2.5%, 97.5%) quantiles of point estimates of the parameters in the propensity score model, $\theta_{X/C}$, and point estimates of τ (maximum likelihood for the traditional approach, posterior-mean estimates for the others). Figure 2 depicts box plots of the widths of 95% uncertainty intervals for estimates of τ , along with coverage probabilities. Confidence intervals for strategies A1 and A2 are obtained from 1000 bootstrap samples for the analysis of each simulated data set. Credible intervals for all other strategies correspond to 95% posterior intervals from the analysis of each simulated data set.

From Figure 1, the failure of Strategy B1 that uses the likelihood in (4) and does not augment the outcome model with covariate adjustment is clear; estimates of $\theta_{X/C}$ do not correspond to the true parameters of the propensity score model, and the resulting point

estimates for τ are clearly biased. This echoes the results of Zigler et al. (2013). In contrast, the Bayesian approach B2 that augments the outcome model to adjust for both $e(C_i; \theta_{X/C})$ and C_i and the two-step Bayesian approach (Strategy C) accurately characterize the treatment assignment mechanism and estimate τ with performance that is comparable to the analogous traditional procedures that sequentially estimate $\hat{e}(C_i; \theta_{X/C})$ and then estimate τ , conditional on $\hat{e}(C_i; \theta_{X/C})$ (Strategies A1 and A2). From Figure 2 note that Strategies A1, A2, and B2 exhibit coverage near the nominal rate, with strategy B2 tending to yield the narrowest uncertainty intervals, which is consistent with known results on variance estimation and propensity score uncertainty (Wooldridge, 2002). The two-step Bayesian Strategy C exhibits the widest intervals and coverage above the nominal rate. The invalid Bayesian approach of Strategy B1 exhibits poor coverage.

5 Conclusion

Incorporating the propensity score into Bayesian procedures for estimating causal effects holds promise for the development of new statistical methodology in modern data applications. Preliminary progress appears in complex latent variable models (Hoshino, 2008), approaches for combining data sources (McCandless et al., 2012), strategies for Bayesian model averaging (Zigler and Dominici, 2014), and approaches for Bayesian estimation of marginal structural models (Saarela et al., 2015). However there exists a fundamental discordance between the need for a fully Bayesian analysis to specify the joint distribution of all observable data (conditional on parameters) and the propensity score's goal of estimating causal effects without modeling the joint distribution of outcome and covariates. Awareness of this discordance is necessary for understanding how Bayesian methods have been and will continue to be used for estimation of causal effects with propensity scores.

This paper considers in detail the setting where inference for causal effects relies on specification of an outcome model that relates $e(C_i; \theta_{X/C})$ to Y_i . An outcome model that adjusts for the propensity score is technically a form of model misspecification of the true data generation, since it does not reflect a probability distribution for $p(Y/X, C)$. Estimating the propensity score in this context presents an interesting analog to an “uncongenial” imputation model (for the propensity score) that does not cohere to a joint distribution with the outcome (Meng, 1994), especially in relation to the two-step Bayesian procedures outlined in Section 3.3. Detailed exploration of this analog is an interesting area of future work. Furthermore, many strategies for estimating causal effects with propensity scores do not specify such an outcome model. These include various matching and stratified estimators, and also inverse probability of treatment weighting (IPTW) procedures that utilize weighted estimating equations to estimate, for example, parameters in a marginal structural model (Robins et al., 2000; Lunceford and Davidian, 2004). These methods do not rely on a likelihood specification. As such, the discussion of Section 3.2 does not directly apply.

However, two-step Bayesian procedures such as that outlined in Section 3.3 hold promise for incorporating Bayesian methods into estimation of causal effects with these non-likelihood-based strategies because τ above can represent any estimator. Upon definition of τ (e.g.,

based on matching or weighting), two-step Bayesian inference could follow from the posterior-predictive distribution of that estimator obtained using expression (6) above, without specifying a model for $p(Y_i|X_i, e_i(C_i, \theta_{X|C}^*), \theta_{Y|X, e^*}))$. Variations of this strategy were employed in Hoshino (2008) and Kaplan and Chen (2012), and this is an important area for continued research.

Acknowledgments

Funding for this work provided by NCI P01 CA134294 and HEI 4909-RFA11-1/12-3. The author thanks Sebastien Haneuse, Francesca Dominici, and Matthew Cefalu for helpful comments and discussion that improved this paper.

References

- An W. Bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*. 2010; 40(1):151–189.
- Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Annals of internal medicine*. 2002; 137(8):693–695. [PubMed: 12379071]
- Brumback BA. A note on using the estimated versus the known propensity score to estimate the average treatment effect. *Statistics & Probability Letters*. 2009; 79(4):537–542.
- Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American journal of epidemiology*. 2003; 158(3):280–287. [PubMed: 12882951]
- Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009; 96(1):187–199.
- D'Agostino RB. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*. 1998; 17(19):2265–2281. [PubMed: 9802183]
- Ho DE, Imai K, King G, Stuart EA. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*. 2007; 15(3):199–236.
- Hoshino T. A Bayesian propensity score adjustment for latent variable modeling and MCMC algorithm. *Computational Statistics & Data Analysis*. 2008; 52(3):1413–1429.
- Imbens GW, Rubin DB. Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. *The Annals of Statistics*. 1997; 25(1):305–327.
- Jin H, Rubin DB. Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*. 2008; 103(481):101–111.
- Kaplan D, Chen J. A Two-Step Bayesian Approach for Propensity Score Analysis: Simulations and Case Study. *Psychometrika*. 2012; 77(3):581–609.
- Liu F, Bayarri MJ, Berger JO. Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*. 2009; 4(1):119–150.
- Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*. 2004; 23(19):2937–2960. [PubMed: 15351954]
- Lunn D, Best N, Spiegelhalter D, Graham G, Neuenschwander B. Combining MCMC with sequential PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics*. 2009; 36(1):19–38. [PubMed: 19132515]
- McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*. 2004; 9(4):403–425. [PubMed: 15598095]
- McCandless LC, Douglas IJ, Evans SJ, Smeeth L. Cutting feedback in Bayesian regression adjustment for the propensity score. *The international journal of biostatistics*. 2010; 6(2):1–22.
- McCandless LC, Gustafson P, Austin PC. Bayesian propensity score analysis for observational data. *Statistics in Medicine*. 2009; 28(1):94–112. [PubMed: 19012268]

- McCandless LC, Richardson S, Best N. Adjustment for Missing Confounders Using External Validation Data and Propensity Scores. *Journal of the American Statistical Association*. 2012; 107(497):40–51.
- Meng XL. Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*. 1994; 9(4):538–558.
- Robins JM, Hernan M, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*. 2000; 11(5):550. [PubMed: 10955408]
- Robins JM, Hernn MA, Wasserman L. Discussion of On Bayesian estimation of marginal structural models. *Biometrics*. 2015 pages n/a–n/a.
- Robins JM, Ritov Y, et al. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in medicine*. 1997; 16(3):285–319. [PubMed: 9004398]
- Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*. 1983; 70(1):41–55.
- Rubin, D. Bayesian statistics. Vol. 2. Elsevier Science Publishers and Valencia University Press; 1985. The use of propensity scores in applied Bayesian inference; p. 463–472.
- Rubin DB. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*. 1978; 6(1):34–58.
- Rubin DB. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*. 2008; 2(3):808–840. Zentralblatt MATH identifier: 1149.62089; Mathematical Reviews number (MathSciNet): MR2516795.
- Rubin DB, Thomas N. Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics*. 1996; 52(1):249–264. ArticleType: research-article / Full publication date: Mar., 1996 / Copyright 1996 International Biometric Society. [PubMed: 8934595]
- Rubin DB, Waterman RP. Estimating the causal effects of marketing interventions using propensity score methodology. *Statistical Science*. 2006:206–222.
- Saarela O, Stephens DA, Moodie EEM, Klein MB. On Bayesian estimation of marginal structural models. *Biometrics*. 2015 pages n/a–n/a.
- Schwartz SL, Li F, Mealli F. A Bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association*. 2011; 106(496):1331–1344.
- Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics*. 2010; 25(1):1–21. [PubMed: 20871802]
- Wooldridge JM. Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*. 2002; 1(2):117–139.
- Zigler CM, Dominici F. Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model-Averaged Causal Effects. *Journal of the American Statistical Association*. 2014; 109(505):95–107. [PubMed: 24696528]
- Zigler CM, Watts K, Yeh RW, Wang Y, Coull BA, Dominici F. Model feedback in Bayesian propensity score estimation. *Biometrics*. 2013; 69(1):263–273. [PubMed: 23379793]

Figure 1.

Mean (squares) and (2.5%, 97.5%) quantiles (lines) of point estimates obtained from analysis of 1000 replicated simulated datasets according to the description in Section 4. Note that estimates θ_{XC} are exactly the same for strategies A1 and A2. Horizontal dashed lines are at the true simulated values.

Figure 2. Uncertainty interval widths for estimates of θ obtained from analysis of 1000 replicated simulated datasets according to the description in Section 4.

Models used to analyze data simulated data. All models are logistic regression, table entries are the linear term on the logit scale. The notation $I(e_i(C_i, \theta_{X|C}))$ denotes the indicator function of whether $e_i(C_i, \theta_{X|C})$ is in the k^{th} quintile of the propensity score distribution.

Table 1

	Model Specification
Propensity Score Model ($m_X(X_i, C_i; \theta_{X C})$)	$\theta_{X C,0} + \sum_{j=1}^6 \theta_{X C,j} C_{ij}$
Outcome Model ($m_Y(Y_i, X_i, e_i(C_i, \theta_{X C}); \theta_{Y X,e})$)	
#1	$\theta_{Y X,e,0} + \theta_{Y X,e,1} X_i + \sum_{k=2}^5 \theta_{Y X,e,k} I(e_i(C_i, \theta_{X C}) \in q_k)$
#2	$\theta_{Y X,e,0} + \theta_{Y X,e,1} X_i + \sum_{k=2}^5 \theta_{Y X,e,k} I(e_i(C_i, \theta_{X C}) \in q_k) + \sum_{j=1}^6 \theta_{Y X,e,5+j} C_{ij}$