Causal Inference: A Missing Data Perspective

Peng Ding and Fan Li

Abstract. Inferring causal effects of treatments is a central goal in many disciplines. The potential outcomes framework is a main statistical approach to causal inference, in which a causal effect is defined as a comparison of the potential outcomes of the same units under different treatment conditions. Because for each unit at most one of the potential outcomes is observed and the rest are missing, causal inference is inherently a missing data problem. Indeed, there is a close analogy in the terminology and the inferential framework between causal inference and missing data. Despite the intrinsic connection between the two subjects, statistical analyses of causal inference and missing data also have marked differences in aims, settings and methods. This article provides a systematic review of causal inference from the missing data perspective. Focusing on ignorable treatment assignment mechanisms, we discuss a wide range of causal inference methods that have analogues in missing data analysis, such as imputation, inverse probability weighting and doubly robust methods. Under each of the three modes of inference—Frequentist, Bayesian and Fisherian randomization—we present the general structure of inference for both finite-sample and super-population estimands, and illustrate via specific examples. We identify open questions to motivate more research to bridge the two fields.

Key words and phrases: Assignment mechanism, ignorability, imputation, missing data mechanism, observational studies, potential outcome, propensity score, randomization, weighting.

1. INTRODUCTION

Causal inference concerns the design and analysis for evaluating the effects of a treatment. A mainstream statistical framework for causal inference is the potential outcomes framework, under which each unit has a set of potential outcomes corresponding to all possible treatment levels (Neyman, 1990, Neyman, 1935, Rubin, 1974, 1977, 1978). Following the dictum "no causation without manipulation" (Rubin, 1975, page 235), a "cause" under the potential outcomes

Peng Ding is Assistant Professor, Department of Statistics, University of California, Berkeley, 425 Evans Hall, Berkeley, California 94720, USA (e-mail: pengdingpku@berkeley.edu). Fan Li is Associate Professor, Department of Statistical Science, Duke University, 122 Old Chemistry Bldg, Durham, North Carolina 27708, USA (e-mail: fl35@duke.edu).

framework strictly refers to a *treatment* or *manipulation*, and a causal effect is defined as a comparison of the potential outcomes under different treatment conditions for the same set of units.

The intrinsic connection to missing data stems from the *fundamental problem of causal inference* (Holland, 1986), that is, for each unit at most one of the potential outcomes—the one corresponding to the treatment to which the unit is exposed—is observed, and the other potential outcomes are missing. Because the potential outcomes of the same unit are never simultaneously observed, the potential outcomes are often referred to as "counterfactuals" in the literature. Therefore, causal inference is inherently a missing data problem and estimating causal effects requires properly handling the missing potential outcomes. Despite the intrinsic connection between the two subjects, statistical analyses of causal inference and missing data—each with a large

and sometimes separate literature—also have marked differences in aims, settings and methods. In this paper, we will provide a systematic review of causal inference from a missing data perspective. Below we first go through some primitives.

1.1 Causal Inference and the Treatment Assignment Mechanism

Consider a simple random sample of units drawn from a target population, indexed by $i \in \{1, \ldots, N\}$, which comprises the participants in a study designed to evaluate the effect of a treatment W on some outcome Y. For example, in comparative effectiveness research, W can be the exposure to a new treatment and Y a health outcome; in economics, W can be the enrollment to a job training program and Y the employment status. Without loss of generality, we consider binary treatments; extension to general treatment regimes has been discussed elsewhere (e.g., Imbens, 2000, Hirano and Imbens, 2004, Imai and van Dyk, 2004). Each unit can potentially be assigned to a treatment w, with $w = 1$ for an active treatment and $w = 0$ for control. Let W_i be the binary variable indicating whether unit i is assigned to the treatment ($W_i = 1$) or to the control ($W_i = 0$). The number of treated and control units are N_1 and N_0, respectively. At baseline, a vector of p pretreatment covariates X_i are observed for unit i. From now on, we use bold font to denote matrices or vectors consisting of the corresponding variables for the N units; for example, let $\mathbf{W} = (W_1, \ldots, W_N)'$ be the N-vector of treatment indicators, and $\mathbf{X} = (X_1', \ldots, X_N')$ be the $N \times p$ covariate matrix. Each unit has a potential outcome under each assignment vector, $Y_i(\mathbf{W})$. Assuming the standard stable unit treatment value assumption (SUTVA) (Rubin, 1980), that is, no interference between units and no different versions of a treatment, each unit has two potential outcomes $Y_i(1)$ and $Y_i(0)$.

The most common causal estimand is the average treatment effect (ATE)—the difference between the average potential outcomes had all units in a target population were taking the treatment versus not. The ATE has both superpopulation (PATE) and finite-sample (SATE) versions:

$$\tau^P \equiv \mathbb{E}\{Y_i(1) - Y_i(0)\},$$
(1)
$$\tau^S \equiv \frac{1}{N} \sum_{i=1}^{N} \{Y_i(1) - Y_i(0)\}.$$

In the PATE estimand τ^P, all potential outcomes are viewed as random variables drawn from a super-population. In the SATE estimand τ^S, all potential outcomes are viewed as fixed values or, equivalently, all inferences are conditional on the vectors of the potential outcomes $\mathbf{Y}(1)$ and $\mathbf{Y}(0)$. SATE has been mostly discussed in the context of randomized experiments, whereas PATE is usually the target estimand in observational studies. The subtle distinction in their definitions leads to important differences in inferential and computational strategies, as discussed later. Other estimands of common interest include the average treatment effects for the treated (ATT) and conditional or individual ATE (Athey and Imbens, 2015, Athey et al., 2017).

Four quantities are associated with each unit i, $\{Y_i(0), Y_i(1), W_i, X_i\}$. Only the potential outcome corresponding to the assigned treatment, $Y_i^{\text{obs}} = Y_i(W_i)$, is observed, and the other potential outcome, $Y_i^{\text{mis}} = Y_i(1 - W_i)$, is missing. Given the observed assignment indicator W_i, there is a one-to-one map between $(Y_i^{\text{obs}}, Y_i^{\text{mis}})$ and $\{Y_i(0), Y_i(1)\}$ with the relationships $Y_i^{\text{obs}} = Y_i(1)W_i + Y_i(0)(1 - W_i)$ and $Y_i^{\text{mis}} = Y_i(1)(1 - W_i) + Y_i(0)W_i$. In general, causal effects are not identifiable without further assumptions. The central identifying assumption concerns the assignment mechanism, that is, the probabilistic process that determines which units receive which treatment condition, and hence which potential outcomes are observed and which are missing. Causal studies can be broadly classified by assignment mechanisms (Imbens and Rubin, 2015, Chapter 3). The vast majority of causal studies assume an *ignorable assignment mechanism*, also known as the *unconfounded* assignment mechanism.

ASSUMPTION 1 (Ignorable assignment mechanism). An assignment mechanism is ignorable (or unconfounded) conditional on $\mathbf{X}$ if it does not depend on the potential outcomes, that is,

$$(2) \qquad \Pr(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = \Pr(\mathbf{W}|\mathbf{X}).$$

Assumption 1 implies that the treatment is randomized among the units with the same value of the observed covariates; in other words, there is no unmeasured confounding. This holds by design in classical randomized experiments where the assignment mechanism is known and depends only on the covariates (Rubin, 1978, Rosenbaum and Rubin, 1983b). However, in observational studies, Assumption 1 cannot be ensured by design or directly validated by the data; it is an untestable sufficient condition allowing for causal inference. In this paper, we require an ignorable assignment for most of the discussion and will comment on its violation in Section 5.

1.2 Missing Data and the Missing Data Mechanism

The assignment mechanism is a special case of a missing data mechanism. Specifically, in the setting of missing or incomplete data, for each unit i, let O_i be the *full data*—the vector containing all the relevant data, observed or unobserved, and R_i be the missing data indicator with the same dimension as O_i. If an entry of O_i is observed, then the corresponding entry of R_i equals 1; if an entry of O_i is missing, then the corresponding entry of R_i equals 0. Let $\mathbf{O} = (O_1', \ldots, O_N')'$ and $\mathbf{R} = (R_1', \ldots, R_N')'$ be the collection of the individual data and missing data indicators, respectively. Let θ and ψ be the set of parameters associated with $\mathbf{O}$ and $\mathbf{R}$, respectively; for simplicity, below we assume that the parameter spaces of θ and ψ are distinct. Given the missing data indicators, we can partition the relevant data into the observed data $\mathbf{O}^{\text{obs}}$ and the missing data $\mathbf{O}^{\text{mis}}$. The missing data mechanism—the conditional distribution of $\mathbf{R}$ given $\mathbf{O}$ and a parameter ψ—is broadly classified into three categories: (i) missing completely at random (MCAR): $\Pr(\mathbf{R}|\mathbf{O}, \theta, \psi) = \Pr(\mathbf{R}|\psi)$; (ii) missing at random (MAR): $\Pr(\mathbf{R}|\mathbf{O}, \theta, \psi) = \Pr(\mathbf{R}|\mathbf{O}^{\text{obs}}, \psi)$; (iii) missing not at random (MNAR) $\Pr(\mathbf{R}|\mathbf{O}, \theta, \psi) = \Pr(\mathbf{R}|\mathbf{O}^{\text{obs}}, \mathbf{O}^{\text{mis}}, \psi)$. MCAR is a special case of MAR, and both are ignorable missing data mechanisms. For more subtle clarification, see Rubin (1976), Seaman et al. (2013) and Mealli and Rubin (2015).

If we consider the full data matrix in a causal setting as $\mathbf{O} = (\mathbf{Y}(1), \mathbf{Y}(0), \mathbf{W}, \mathbf{X})$, then the missing data indicator is $\mathbf{R} = (\mathbf{W}, \mathbf{1}_N - \mathbf{W}, \mathbf{1}_N, \mathbf{1}_{N \times p})$; consequently, the observed data contain $\mathbf{O}^{\text{obs}} = (\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X})$ and the missing data contain $\mathbf{O}^{\text{mis}} = \mathbf{Y}^{\text{mis}}$. The fundamental problem of causal inference is reflected by the fact that $\mathbf{1}_N - \mathbf{W}$, the missing data indicators for $\mathbf{Y}(1)$, and $\mathbf{W}$, the missing data indicators for $\mathbf{Y}(0)$, sum to $\mathbf{1}_N$; that is, the two potential outcomes cannot be simultaneously observed, and thus missing data are unavoidable. Gelman et al. [(2014), page 198] call these missing potential outcomes *intentional* missing data, in contrast to the *unintentional* missing data due to unfortunate circumstances such as survey nonresponse, loss to follow-up, censored measurements or clerical errors. Unintentional missing data are extensively discussed in other articles in this issue; in this article, we focus on intentional missing data in causal inference, assuming $\mathbf{W}$ and $\mathbf{X}$ are fully observed.

1.3 Connection and Distinction

Based on the representation in Section 1.2, the assignment mechanism creates a missing data mechanism of the potential outcomes. There is a broad parallel between the classification of assignment mechanisms in causal inference and the classification of missing data mechanisms. Specifically, the assignment mechanism of completely randomized experiments, namely, $\Pr(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \theta, \psi) = \Pr(\mathbf{W}|\psi)$, is in parallel to MCAR. The assignment mechanism of observational studies with unmeasured confounding, namely, $\Pr(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \theta, \psi) = \Pr(\mathbf{W}|\mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}}, \mathbf{X}, \psi)$, is in parallel to MNAR. MAR generally corresponds to ignorable assignment mechanisms. However, because the observed covariates, treatment and outcomes play very different roles in causal inference, the original form of ignorable assignment mechanism can be generalized to a class of assignment mechanisms that satisfy the condition: $\Pr(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = \Pr(\mathbf{W}|\mathbf{Y}^{\text{obs}}, \mathbf{X})$. For example, this class includes stratified randomized experiments, the assignment mechanism of which only depends on the covariates, namely, $\Pr(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \theta, \psi) = \Pr(\mathbf{W}|\mathbf{X}, \psi)$. Another example is the randomized experiments where the units $i = 1, \ldots, N$ are assigned sequentially, with the assignment of unit i depending only on the covariates of units $1, \ldots, i$ and the observed outcomes of units $1, \ldots, i-1$, that is, $\Pr(W_i|W_1, \ldots, W_{i-1}, \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \theta, \psi) = \Pr(W_i|Y_1^{\text{obs}}, \ldots, Y_{i-1}^{\text{obs}}, \mathbf{X}, \psi)$ (Imbens and Rubin, 2015, Chapter 3). A third example is the sequentially ignoble assignment mechanism in studies with time-varying treatments, where the treatment assignment is ignorable at each time point t conditional on the observed history until time $t-1$, including baseline covariates, observed treatments and immediate outcomes (i.e., time-varying covariates) (Robins, 1986). Evident from the case of MAR, there is a richer class of assignment mechanisms than missing data mechanisms. In fact, there are a variety of complex assignment mechanisms with no analogous missing data mechanisms, such as latent ignorable assignment with intermediate variables (Frangakis and Rubin, 1999) and locally ignorable assignment in regression discontinuity designs (Li, Mattei and Mealli, 2015).

There are several important distinctions between causal inference and the standard missing data analysis. First, the central goal of causal inference is (unconfounded) comparison—comparing the potential outcomes of the same units under two or more treatment conditions, whereas the common goal in standard missing data analysis is to infer a population parameter—not necessarily comparison—defined on the full data, $\theta = f(\mathbf{O})$, from the observed data $\mathbf{O}^{\text{obs}}$. This has at

least two consequences: (a) The crucial concept of co-variate balance between two groups in causal inference is usually not considered in missing data analysis; see Zubizarreta (2015) for an exception; (b) The widely used Fisher randomization test for causal comparisons has no analog in missing data analysis; see Section 5.1 for details. Second, the missingness of potential outcomes is highly structured: for each unit, the missing data indictors of all potential outcomes must sum to one, and thus at least half of the potential outcomes are missing by design. Critically, there is no information to infer the association between the two potential outcomes. In contrast, in standard incomplete data setting, missing data can occur in any part of the full data matrix without structural constraints and usually there is information to infer the association between any two variables. Third, causal inference differentiates between pre-treatment and post-treatment variables, whereas the standard missing data methods do not differentiate between them. For example, it is usually scientifically meaningless to postulate a model of pretreatment variable conditional on post-treatment variables, for example, $\Pr(\mathbf{X}|\mathbf{Y}(1), \mathbf{Y}(0))$, but such ordering restriction between variables is rarely imposed in missing data analysis. For example, ordering information between longitudinal variables is usually ignored in the popular multiple imputation by chained equation (MICE) algorithm (van Buuren, 2012).

Despite these distinctions, the inferential framework of causal inference and missing data analysis are closely related, with common roots in survey sampling. In particular, two overarching methods underpin both causal inference and missing data analysis: weighting and imputation. Although mathematically imputation can be represented as a special form of weighting, these two methods are often derived from different perspectives and implemented differently. Weighting methods weight—usually based on the probability of being missing or assigned to one group—the observed data to represent the full data, and imputation methods physically impute the missing values—often based on a stochastic model—using the observed data. Moreover, sensitivity analysis on the missing data mechanism and the treatment assignment mechanism is routinely conducted in both domains.

Focusing on ignorable assignment mechanisms, in this article we review a wide range of causal inference methods that have analogues in missing data analysis, such as imputation, inverse probability weighting and doubly robust methods. We also provide examples where ideas originated from causal inference lend to missing data analysis, such as truncation by death and covariate balance. We organize the review by the mode of inference with Frequentist in Section 2, Bayesian in Section 3 and Fisherian in Section 4. Within each mode of inference, we first present the general structure of causal inference and then illustrate via specific examples; also, depending on the study setting, we switch between two sampling models of the units: we usually, with a few exceptions, adopt the finite-sample model for randomized experiments and the super-population model for observational studies. Section 5 concludes with a discussion on open questions and future research directions.

2. THE FREQUENTIST PERSPECTIVE

In a broad sense, the Frequentist or classical inference focuses on repeated sampling evaluation of statistical procedures, such as unbiasedness or consistency, variance and efficiency, and mean squared error of a point estimator, coverage rate of an interval estimator, optimality and minimaxity. Here, we discuss the frequentist perspective in a narrow sense, reviewing the procedures motivated by frequency properties. In particular, we first connect the classical randomization-based inference and model-based imputation in completely randomized experiments, we then review methods using weighting or generally unbiased estimating equations in unconfounded observational studies and we finally discuss principal stratification in the case of post-treatment variables.

2.1 Model-Based Imputation in Randomized Experiments

For illustration purposes, we consider a classical completely randomized experiment with covariates, where the inferential goal is to find point and interval estimators for the SATE, τ^S. Recall that in finite-sample inference all potential outcomes $\mathbf{Y}(1)$ and $\mathbf{Y}(0)$ are viewed as fixed and the randomness comes solely from the treatment indicators $\mathbf{W}$. It is intuitive to impute the missing values—$Y_i(0)$ for the treated units and $Y_i(1)$ for the control units—based on some models of the potential outcomes, and use the imputed potential outcomes, denoted by $\hat{Y}_i(w)$, to estimate the SATE, giving the following *predictive* estimator:

$$
\begin{aligned}
\hat{\tau}^{\text{pre}} = \frac{1}{N}\Bigg\{ &\sum_{i=1}^{N} W_i Y_i(1) + \sum_{i=1}^{N}(1 - W_i)\hat{Y}_i(1) \\
&- \sum_{i=1}^{N} W_i \hat{Y}_i(0) - \sum_{i=1}^{N}(1 - W_i)Y_i(0) \Bigg\}.
\end{aligned}
\tag{3}
$$

As a starting point, we consider fitting linear models based on least squares: in the treatment group, $\widehat{Y}_i(1) = X_i'\widehat{\beta}_1 + \widehat{\gamma}_1$, and in the control group, $\widehat{Y}_i(0) = X_i'\widehat{\beta}_0 + \widehat{\gamma}_0$. Here, we highlight that the regression lines include intercepts, and the covariates are centered at zero (i.e., $\bar{X} = \sum_{i=1}^N X_i = 0$). Define $(\bar{Y}_1^{\text{obs}}, \bar{Y}_0^{\text{obs}})$ and $(\bar{X}_1^{\text{obs}}, \bar{X}_0^{\text{obs}})$ as the means of the outcomes and covariates under treatment and control groups, respectively. Because the least squares coefficients satisfy $\widehat{\gamma}_w = \bar{Y}_w^{\text{obs}} - \widehat{\beta}_w' \bar{X}_w^{\text{obs}}$ ($w = 1, 0$), using some algebra we can simplify the estimator in (3) to

$$
(4) \quad
\begin{aligned}
\widehat{\tau}^{\text{pre}} &= \widehat{\gamma}_1 - \widehat{\gamma}_0 \\
&= (\bar{Y}_1^{\text{obs}} - \widehat{\beta}_1' \bar{X}_1^{\text{obs}}) - (\bar{Y}_0^{\text{obs}} - \widehat{\beta}_0' \bar{X}_0^{\text{obs}}).
\end{aligned}
$$

Therefore, $\widehat{\tau}^{\text{pre}}$ is identical to the regression adjusted estimator in Lin (2013) motivated by an estimation strategy in survey sampling (Cochran, 2007). Without using any covariates or simply setting $\widehat{\beta}_1 = \widehat{\beta}_0 = 0$, the estimator $\widehat{\tau}^{\text{pre}}$ reduces to the classical Neymanian unbiased difference-in-means estimator, $\widehat{\tau}^{\text{Neyman}} = \bar{Y}_1^{\text{obs}} - \bar{Y}_0^{\text{obs}}$, which essentially imputes the missing potential outcomes using the observed sample means $\bar{Y}_1^{\text{obs}}$ and $\bar{Y}_0^{\text{obs}}$. Lin (2013) studied the repeated sampling properties of $\widehat{\tau}^{\text{pre}}$ over the distribution of $\mathbf{W}$, showing that (a) it is consistent and at least as efficient as $\widehat{\tau}^{\text{Neyman}}$ even if the linear potential outcome models are *misspecified*, (b) it is identical to the regression coefficient of $\mathbf{W}$ in the linear regression $\mathbf{Y}^{\text{obs}} \sim \mathbf{W} + \mathbf{X} + \mathbf{W} \times \mathbf{X}$ with full interactions between treatment and covariates, and (c) asymptotically the Huber–White variance estimator (Huber, 1967, White, 1980) is conservative for the randomization-based sampling variance of $\widehat{\tau}^{\text{pre}}$. Li and Ding [(2017), Example 9] supplemented these results with the asymptotic normality and the optimality of $\widehat{\tau}^{\text{pre}}$.

To emphasize, the estimator $\widehat{\tau}^{\text{pre}}$ is motivated by imputing missing potential outcomes based on linear models, but its frequency properties over complete randomization do not require any modeling assumptions. Recently, Bloniarz et al. (2016) extended the above approach to deal with high dimensional covariates, replacing the least squares coefficients by the LASSO coefficients (Tibshirani, 1996). Invoking super-population assumptions, Wager et al. (2016) considered other machine learning methods as generalization of the least squares method. So far all these discussions are in the context of completely randomized experiments; rigorous investigations of other types of experiments are desirable.

2.2 Imputation and Weighting in Observational Studies

As the PATE, τ^P, is the primary target estimand in most observational studies, we now shift to the super-population view. We assume the units are a simple random sample from the target population, that is, $\{Y_i(1), Y_i(0), W_i, X_i\}_{i=1}^N$ are independent and identically distributed. Then a sufficient condition for identifying τ^P is strong ignorability (Rosenbaum and Rubin, 1983b), consisting of Assumption 1 and the overlap (also known as positivity) assumption, which requires the *propensity score*, $e(X_i) \equiv \Pr(W_i = 1 | X_i)$, is strictly between 0 and 1 for all values of X_i.

2.2.1 *Imputation methods.* The most popular method for causal inference in observational studies is regression adjustment (Rubin, 1979), which, similar to the case in randomized experiments, essentially imputes missing potential outcomes from a regression model. Let (p_1, p_0) be the proportions of treated and control units. Under Assumption 1, the conditional mean functions of the potential outcomes satisfy that for any w, w', x,

$$
(5) \quad
\begin{aligned}
m_w(x) &\equiv \mathbb{E}\{Y_i(w) | X_i = x\} \\
&= \mathbb{E}\{Y_i(w) | W_i = w', X_i = x\} \\
&= \mathbb{E}\{Y_i^{\text{obs}} | W_i = w, X_i = x\},
\end{aligned}
$$

where the first identity is the definition, the second identity is about the counterfactual mean if $w \neq w'$ and the last identify is the conditional mean function of the observed outcomes. Therefore, the PATE can be identified from the observed data:

$$
(6) \quad
\begin{aligned}
\tau^P &= \big[p_1 \mathbb{E}\{Y_i(1) | W_i = 1\} + p_0 \mathbb{E}\{Y_i(1) | W_i = 0\} \big] \\
&\quad - \big[p_1 \mathbb{E}\{Y_i(0) | W_i = 1\} + p_0 \mathbb{E}\{Y_i(0) | W_i = 0\} \big] \\
&= \Big[p_1 \mathbb{E}\{Y_i^{\text{obs}} | W_i = 1\} + p_0 \int m_1(X) F_0(\mathrm{d}x) \Big] \\
&\quad - \Big[p_1 \int m_0(x) F_1(\mathrm{d}x) + p_0 \mathbb{E}\{Y_i^{\text{obs}} | W_i = 0\} \Big],
\end{aligned}
$$

where $F_1(x)$ and $F_0(x)$ are the distributions of the covariates in the treatment and control group, respectively. If $\{\widehat{m}_1(x), \widehat{m}_0(x)\}$ are the fitted conditional mean functions based on the treated and control outcome data, then $\{\widehat{Y}_i(1), \widehat{Y}_i(0)\} = \{\widehat{m}_1(X_i), \widehat{m}_0(X_i)\}$ are the fitted values of the possibly missing potential outcomes. This strategy gives the predictive estimator (3), which reduces to the regression-adjusted estimator (4) if both conditional mean functions are modeled as linear.

The predictive estimator (3) also reduces to the popular matching-without-replacement estimator if both $\widehat{m}_1(X_i)$ and $\widehat{m}_0(X_i)$ are fitted by the nearest-neighbor regression (Abadie and Imbens, 2006). It is worth commenting more on matching, which has a long tradition in statistics and particularly in causal inference. An early example of matching was given in Chapin (1947). Championed by Rubin (a collection of important papers were reprinted in Rubin, 2006b) and later by Rosenbaum (Rosenbaum, 2010, Part II), matching has been increasingly embraced by both methodologists and practitioners as a model-free alternative to the regression adjustment method. Instead of a regression model, matching methods use the mean outcome of the units with similar covariates in the opposite group (i.e., matches) to impute the missing potential outcome. In standard missing data analysis, matching-based imputation is also commonly used, where the missing values of one unit are imputed by the observed values of the units who are matched on other characteristics. However, in large complex datasets with incomplete data, model-based imputation methods are generally more flexible and versatile than matching-based ones (e.g., van Buuren, 2012, Li et al., 2014, Murray and Reiter, 2016). Indeed, in these cases matching is often combined with regression models, such as the predictive mean matching method (Rubin, 1986, Little, 1988). The idea of combining matching and regression is also used in causal inference, for example, in the bias-corrected matching estimator (Abadie and Imbens, 2011). In this article, we view matching as a special case of imputation methods and refer interested readers to elsewhere for a comprehensive review of the vast literature on matching (e.g., Stuart, 2010, Rosenbaum, 2010).

Another representation of the PATE is

$$
(7) \quad \begin{aligned} \tau^P &= \int \{m_1(x) - m_0(x)\} F(\mathrm{d}x) \\ &= \mathbb{E}\{m_1(X) - m_0(X)\}, \end{aligned}
$$

where $F(x)$ is the distribution of the covariates in the whole population. This representation motivates the following *projective* estimator:

$$
\widehat{\tau}^{\mathrm{pro}} = \frac{1}{N} \sum_{i=1}^{N} \{\widehat{m}_1(X_i) - \widehat{m}_0(X_i)\},
$$

which is in general different from the *predictive* estimator $\widehat{\tau}^{\mathrm{pre}}$, as discussed in the survey sampling literature (Firth and Bennett, 1998). However, the Normal equation of the linear model and the score equation of the logistic model both satisfy $\sum_{W_i=w} Y_i^{\mathrm{obs}} =$ $\sum_{W_i=w} \widehat{m}_w(X_i)$, ensuring that $\widehat{\tau}^{\mathrm{pro}} = \widehat{\tau}^{\mathrm{pre}}$ if we fit separate linear or logistic models with intercepts for the observed outcomes under treatment and control. In contrast to the case of completely randomized experiments as discussed in Section 2.1, consistency of the estimators $\widehat{\tau}^{\mathrm{pro}}$ and $\widehat{\tau}^{\mathrm{pre}}$ in observational studies does rely on correct specification of the conditional mean functions $m_1(x)$ and $m_0(x)$ (Tsiatis et al., 2008).

A special case of the regression adjustment/ imputation method is to use the estimated propensity scores as a predictor additional to the covariates in the outcome model; such a model has been shown in many empirical studies to outperform the same model without the propensity scores. In particular, Little and An (2004) and Zhang and Little (2009) advocated using penalized splines of the estimated propensity scores in the outcome model, in the context of both causal inference and missing data. The common idea originates from survey methodology, where the sampling probabilities (equivalently the survey weights) usually contain information about the units besides the covariates and have been used to augment the inference about population parameters.

2.2.2 *Weighting methods.* A second class of widely used causal inference methods in observational studies is weighting, particularly propensity score weighting. Rosenbaum and Rubin (1983b) showed that if the assignment mechanism is unconfounded given the covariates X_i (i.e., Assumption 1), then it is also unconfounded given the scalar propensity score $e(X_i)$. The classical Horvitz–Thompson estimator (Horvitz and Thompson, 1952) in survey sampling motivates the following representation of the PATE:

$$
(8) \quad \tau^P = \mathbb{E}\left\{\frac{W_i Y_i^{\mathrm{obs}}}{e(X_i)}\right\} - \mathbb{E}\left\{\frac{(1 - W_i) Y_i^{\mathrm{obs}}}{1 - e(X_i)}\right\}.
$$

This suggests that one can define an inverse probability weight (IPW) for each unit in the treatment as $1/e(X_i)$ and in the control group as $1/\{1 - e(X_i)\}$, and estimate the PATE by the difference in the weighted average of the outcomes in the two groups. Li, Morgan and Zaslavsky (2018) show that the IPW is a special case of the general class of *balancing weights*, which balance the weighted distributions of the covariates between treatment and control groups in any pre-specified target population.

If we fit a model for $\Pr(W_i = 1 | X_i)$ and obtain the estimated propensity score $\widehat{e}(X_i)$, then the sample analogue of (8) gives the IPW estimator of the PATE:

$$
(9) \quad \widehat{\tau}^{\mathrm{ipw}} = \frac{1}{N} \sum_{i=1}^{N} \frac{W_i Y_i^{\mathrm{obs}}}{\widehat{e}(X_i)} - \frac{1}{N} \sum_{i=1}^{N} \frac{(1 - W_i) Y_i^{\mathrm{obs}}}{1 - \widehat{e}(X_i)}.
$$

The consistency of $\widehat{\tau}^{\text{ipw}}$ requires a correctly specified propensity score model. If we replace the N's in the denominators by $\sum_{i=1}^{N} W_i/\widehat{e}(X_i)$ and $\sum_{i=1}^{N}(1 - W_i)/\{1 - \widehat{e}(X_i)\}$, respectively, the resulting estimator is called the Hájek estimator, which often has smaller mean squared error than estimator (9) (Hájek, 1971). When the fitted values $\widehat{e}(X_i)$ are close to 0 or 1, the IPW estimator is unstable, leading to large variance and bias. Common remedies include trimming units with extreme weights (Crump et al., 2009, Yang and Ding, 2018) and coarsening propensity scores (Zhou et al., 2015). Another approach is to stratify on $\widehat{e}(X_i)$ (Rosenbaum and Rubin, 1983b), which can be viewed as a special case of $\widehat{\tau}^{\text{ipw}}$ with the estimated propensity scores coarsened before being plugged into (9). Li, Morgan and Zaslavsky (2018) advocated changing the inverse probability weights in $\widehat{\tau}^{\text{ipw}}$ to the overlap weights [i.e., weights control units by $e(X_i)$ and treated units by $1 - e(X_i)$], which shifts the focus to the target population with the most overlap and automatically bypasses the problem of extreme weights.

The Horvitz–Thompson idea that a unit being weighted inverse-proportionally to its probability of being sampled applies to both causal inference and missing data. Unsurprisingly, the IPW approach is also widely adopted in missing data analysis, particularly for handling incomplete outcome data. For example, in order to estimate a population parameter from a sample with incomplete outcome data, we can first estimate the probability of being observed based on the covariates for each unit, corresponding to the propensity score, and then run a weighted regression model where each observed unit is inversely weighted by its probability of being observed.

2.2.3 *Combining imputation and weighting*: *Doubly robust methods*. A third class of popular methods is the doubly robust estimation, which combines the virtues of regression adjustment and propensity score weighting. Based on (7) and (8), it is straightforward to verify that

$$
\begin{aligned}
\tau^P = {}& \mathbb{E}\left\{ \frac{W_i Y_i^{\text{obs}}}{e(X_i)} - \frac{W_i - e(X_i)}{e(X_i)} m_1(X_i) \right\} \\
& - \mathbb{E}\left\{ \frac{(1 - W_i)Y_i^{\text{obs}}}{1 - e(X_i)} + \frac{W_i - e(X_i)}{1 - e(X_i)} m_0(X_i) \right\} \\
= {}& \mathbb{E}\left[m_1(X_i) + \frac{W_i\{Y_i^{\text{obs}} - m_1(X_i)\}}{e(X_i)} \right] \\
& - \mathbb{E}\left[m_0(X_i) + \frac{(1 - W_i)\{Y_i^{\text{obs}} - m_0(X_i)\}}{1 - e(X_i)} \right].
\end{aligned}
$$

This motivates the following estimator:

$$
\begin{aligned}
\widehat{\tau}^{\text{dr}} = {}& \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{W_i Y_i^{\text{obs}}}{\widehat{e}(X_i)} - \frac{W_i - \widehat{e}(X_i)}{\widehat{e}(X_i)} \widehat{m}_1(X_i) \right\} \\
& - \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{(1 - W_i)Y_i^{\text{obs}}}{1 - \widehat{e}(X_i)} \right. \\
& \qquad\qquad \left. + \frac{W_i - \widehat{e}(X_i)}{1 - \widehat{e}(X_i)} \widehat{m}_0(X_i) \right\} \\
= {}& \frac{1}{N} \sum_{i=1}^{N} \left[\widehat{m}_1(X_i) + \frac{W_i\{Y_i^{\text{obs}} - \widehat{m}_1(X_i)\}}{\widehat{e}(X_i)} \right] \\
& - \frac{1}{N} \sum_{i=1}^{N} \left[\widehat{m}_0(X_i) \right. \\
& \qquad\qquad \left. + \frac{(1 - W_i)\{Y_i^{\text{obs}} - \widehat{m}_0(X_i)\}}{1 - \widehat{e}(X_i)} \right].
\end{aligned}
$$

(10)

(11)

The equivalent forms (10) and (11) follow from trivial mathematical manipulations but have profound statistical implications. We can view (10) as the IPW estimator augmented by an outcome model, and show that if the propensity score model is correctly specified then $\widehat{\tau}^{\text{dr}}$ is consistent for τ no matter whether the outcome models is correctly specified. Alternatively, we can view (11) as the projective estimator augmented by inverse probability weighting, and show that if the outcome models are correctly specified then $\widehat{\tau}^{\text{dr}}$ is consistent for τ no matter whether the propensity score model is correctly specified. This property is called double-robustness (DR) (Scharfstein, Rotnitzky and Robins, 1999, Bang and Robins, 2005), that is, $\widehat{\tau}^{\text{dr}}$ is consistent for τ if either the propensity score model or the outcome models are correctly specified. Moreover, the DR estimator achieves the semiparametric efficiency bound if both the propensity score and outcome models are correctly specified (Robins, Rotnitzky and Zhao, 1995, Robins and Ritov, 1997). For more discussion on efficiency, see Hahn (1998), Hirano, Imbens and Ridder (2003), Imbens (2004) and Qin (2017). The DR estimator is closely connected to missing data analysis; in fact, the concept of DR was first proposed in the context of missing data and was later formally extended to causal inference (Lunceford and Davidian, 2004, Bang and Robins, 2005).

The DR estimator has attracted increasing attention from both theoreticians and practitioners. Below we comment on several issues appearing frequently in the recent literature. First, the propensity score model is

often chosen to be logistic in practice, and the outcome models linear. With low dimensional covariates, we can add power series of the covariates to approximate complex propensity score and outcome models (Newey, 1997, Hirano, Imbens and Ridder, 2003, Mercatanti and Li, 2014). Second, with high dimensional covariates, the DR estimator works well under more general assumptions than many other estimators, and in order to achieve better asymptotic properties, for example, $\sqrt{n}$-consistency, the DR estimator is probably the only way to go (van der Laan and Rose, 2011, Belloni, Chernozhukov and Hansen, 2014, Chernozhukov et al., 2016, Athey, Imbens and Wager, 2018, Belloni et al., 2017). More importantly, in high dimensions, it is crucial to regularize the estimators as suggested by, for example, Chernozhukov et al. (2016) and Athey et al. (2017). Third, the unstable properties of $\hat{\tau}^{\mathrm{dr}}$ can be more severe than $\hat{\tau}^{\mathrm{ipw}}$ with extreme estimated propensity scores, especially when both the propensity score and outcomes models are misspecified (Kang and Schafer, 2007). In this case, trimming or truncating the estimated propensity scores seems crucial for reliable analysis. Alternatively, Graham, de Xavier Pinto and Egel (2012), Hainmueller (2012) and Imai and Ratkovic (2014) proposed to construct weights that balance the covariates directly, which can avoid the problem of extreme estimated propensity scores. Interestingly, the idea of building propensity score models based on covariate balance has long been used in the causal inference, but it is only recently adopted in missing data problems. Specifically, Zubizarreta (2015) constructed more stable weighting estimators for missing data analysis by directly constructing weights to balance covariates of units with observed and unobserved outcomes.

2.3 Principal Stratification

An important problem in causal inference is to deal with post-treatment variables—variables that are potentially affected by the treatment and possibly affect the outcome. Cochran (1957) and Rosenbaum (1984b) show that adjusting for post-treatment variables in the same fashion as adjusting for pre-treatment covariates would bias the estimation of causal effects in general. Since the landmark papers by Angrist, Imbens and Rubin (1996) and Frangakis and Rubin (2002), a large literature on this topic has been developed. The post-treatment variable settings include a wide range of specific examples, such as treatment noncompliance (e.g., Imbens and Rubin, 1997), outcomes truncated by "death" (e.g., Rubin, 2006a, Zhang, Rubin and Mealli,

2009), surrogate endpoints (e.g., Gilbert and Hudgens, 2008, Zigler and Belin, 2012), mediation analysis (e.g., VanderWeele, 2008, Gallop et al., 2009, Elliott, Raghunathan and Li, 2010) and fuzzy regression discontinuity designs (e.g., Li, Mattei and Mealli, 2015, Chib and Jacobi, 2016).

2.3.1 *Randomized experiments with noncompliance.* To introduce the basic setup, we start with the simplest setting of a completely randomized experiment with noncompliance. We need some new notation to accommodate the post-treatment setting. For unit i ($i = 1, \dots, N$) in a study sample, let Z_i be the treatment assigned to (1 for treatment and 0 for control), and W_i^{obs} be the treatment received (1 for treatment and 0 for control). When $Z_i \neq W_i^{\mathrm{obs}}$, noncompliance occurs. Because W_i is a post-treatment variable, it has two potential outcomes, $W_i(0)$ and $W_i(1)$, with $W_i^{\mathrm{obs}} = W_i(Z_i)$. As before, the outcome Y_i also has two potential outcomes, $Y_i(0)$ and $Y_i(1)$. Angrist, Imbens and Rubin (1996) classified the units according to the joint potential treatment statuses $U_i = (W_i(1), W_i(0))$, which was later called the principal stratification by Frangakis and Rubin (2002). The principal stratum U_i takes four values: $(1, 0) = \mathrm{c}$ for compliers, $(0, 0) = \mathrm{n}$ for never takers, and $(1, 1) = \mathrm{a}$ for always takers and $(0, 1) = \mathrm{d}$ for defiers. Due to the fundamental problem of causal inference, individual principal stratum U_i are not observed.

The key property of principal strata is that they are, by definition, not affected by the treatment assignment, and thus can be regarded as a pre-treatment variable. Therefore, comparisons of $Y_i(1)$ and $Y_i(0)$ within a principal stratum—the principal causal effects (PCEs)—have a causal interpretation:

$$
\begin{aligned}
\tau_u &= \mathbb{E}\{Y_i(1) - Y_i(0) | U_i = u\} \\
&= \mathbb{E}\{Y_i(1) | U_i = u\} - \mathbb{E}\{Y_i(0) | U_i = u\}, \\
&\quad (u = \mathrm{n}, \mathrm{c}, \mathrm{a}, \mathrm{d}).
\end{aligned}
$$

Note that PCE also has both finite-sample and super-population versions. The conventional causal estimand in the noncompliance setting is the intention-to-treat (ITT) effect, ignoring noncompliance; ITT can be decomposed into the sum of the four PCEs:

$$
\tau^P = \mathbb{E}\{Y_i(1) - Y_i(0)\} = \sum_{u=\mathrm{c},\mathrm{n},\mathrm{a},\mathrm{d}} \pi_u \tau_u^P,
$$

where $\pi_u = \Pr(U_i = u)$ is the proportion of the stratum u. Inferences about the PCEs are often scientifically relevant; for example, the PCE for compliers—commonly known as the complier average causal effect

(CACE) or the local average treatment effect (LATE) (Imbens and Angrist, 1994)—is informative about the efficacy of the treatment received, whereas the ITT is informative about the effectiveness of the treatment assigned.

The main challenge to the inference of principal stratification is the missing latent stratum membership U_i. Specifically, the observed cells of Z and W^{obs} usually consists of a mixture of units from more than one principal strata, and one has to disentangle the causal effects for different principal strata from observed data. Therefore, the general structure of inference with principal stratification is similar to that of a mixture model, and standard methods for mixture models such as the EM algorithm (Dempster, Laird and Rubin, 1977) and data augmentation (Tanner and Wong, 1987) are routinely used in principal stratification, as illustrated below in Section 2.3.2.

2.3.2 *Implicit weighting via the EM algorithm.* Throughout the discussion, we will use $\Pr(\cdot|\cdot)$ and $\theta_{\cdot|\cdot}$ to denote generic conditional distributions and the corresponding parameters, respectively. If we have full parametric models for $\Pr(U_i|X_i, \theta_{W|X})$ and $\Pr(Y_i(z)|U_i, X_i, \theta_{Y|U,X})$, then we can use the EM algorithm to obtain the maximum likelihood estimator for $\theta = (\theta_{W|X}, \theta_{Y|U,X})$ treating $\mathbf{U}$ as the missing data. Introducing the shorthand notation $\pi_{i,u} = \Pr(U_i = u|X_i, \theta_{W|X})$ and $f_{i,uz} = \Pr(Y_i(z)|U_i = u, X_i, \theta_{Y|U,X})$, the complete-data log-likelihood, based on $(\mathbf{Z}, \mathbf{W}^{\mathrm{obs}}, \mathbf{U}, \mathbf{Y}^{\mathrm{obs}}, \mathbf{X})$, is

$$
\begin{aligned}
& l(\theta, \mathbf{U}) \\
& = \sum_{Z_i=1, W_i^{\mathrm{obs}}=1} I(U_i = \mathrm{c})\{\log \pi_{i,\mathrm{c}} + \log f_{i,\mathrm{c}1}\} \\
& \quad + \sum_{Z_i=1, W_i^{\mathrm{obs}}=1} I(U_i = \mathrm{a})\{\log \pi_{i,\mathrm{a}} + \log f_{i,\mathrm{a}1}\} \\
& \quad + \sum_{Z_i=1, W_i^{\mathrm{obs}}=0} I(U_i = \mathrm{n})\{\log \pi_{i,\mathrm{n}} + \log f_{i,\mathrm{n}1}\} \\
& \quad + \sum_{Z_i=1, W_i^{\mathrm{obs}}=0} I(U_i = \mathrm{d})\{\log \pi_{i,\mathrm{d}} + \log f_{i,\mathrm{d}1}\} \\
& \quad + \sum_{Z_i=0, W_i^{\mathrm{obs}}=1} I(U_i = \mathrm{a})\{\log \pi_{i,\mathrm{a}} + \log f_{i,\mathrm{n}0}\} \\
& \quad + \sum_{Z_i=0, W_i^{\mathrm{obs}}=1} I(U_i = \mathrm{d})\{\log \pi_{i,\mathrm{d}} + \log f_{i,\mathrm{d}0}\}
\end{aligned}
$$

$$
\begin{aligned}
& \quad + \sum_{Z_i=0, W_i^{\mathrm{obs}}=0} I(U_i = \mathrm{c})\{\log \pi_{i,\mathrm{c}} + \log f_{i,\mathrm{c}0}\} \\
& \quad + \sum_{Z_i=0, W_i^{\mathrm{obs}}=0} I(U_i = \mathrm{n})\{\log \pi_{i,\mathrm{n}} + \log f_{i,\mathrm{n}0}\}.
\end{aligned}
$$

We emphasize that $l(\theta, \mathbf{U})$ is a function of the parameter θ and depends on the missing principal strata $\mathbf{U}$.

In the E-step of the EM algorithm, we need to find $Q(\theta|\theta^{[t]}) = \mathbb{E}\{l(\theta, \mathbf{U})|\mathbf{Z}, \mathbf{W}^{\mathrm{obs}}, \mathbf{Y}^{\mathrm{obs}}, \mathbf{X}, \theta^{[t]}\}$, the conditional expectation of the complete-data log-likelihood given the observed data and the value of the parameter at iteration t, which reduces to calculating the conditional probabilities of $\Pr^{[t]}(U_i = u) = \Pr(U_i = u|Z_i, W_i^{\mathrm{obs}}, Y_i^{\mathrm{obs}}, X_i, \theta^{[t]})$. For example, for unit with $Z_i = 1$ and $W_i^{\mathrm{obs}} = 1$, we have

$$
\Pr^{[t]}(U_i = \mathrm{c}) = \frac{\pi_{i,\mathrm{c}}^{[t]} f_{i,\mathrm{c}1}^{[t]}}{\pi_{i,\mathrm{c}}^{[t]} f_{i,\mathrm{c}1}^{[t]} + \pi_{i,\mathrm{a}}^{[t]} f_{i,\mathrm{a}1}^{[t]}},
$$

$$
\Pr^{[t]}(U_i = \mathrm{a}) = 1 - \Pr^{[t]}(U_i = \mathrm{c}),
$$

where $\pi_{i,u}^{[t]}$ and $f_{i,u1}^{[t]}$ are evaluated at $\theta^{[t]}$. This effectively creates two weighted observations: one has $(Z_i = 1, W_i^{\mathrm{obs}} = 1, U_i = \mathrm{c}, Y_i^{\mathrm{obs}}, X_i)$ with weight $\Pr^{[t]}(U_i = \mathrm{c})$, and the other has $(Z_i = 1, W_i^{\mathrm{obs}} = 1, U_i = \mathrm{a}, Y_i^{\mathrm{obs}}, X_i)$ with weight $\Pr^{[t]}(U_i = \mathrm{a})$. For units with other combinations of Z_i and W_i^{obs}, we can similarly obtain $\Pr^{[t]}(U_i = u)$ and create weighted observations.

In the M-step of the the EM algorithm, we need to maximize $Q(\theta|\theta^{[t]})$, or, equivalently, the log-likelihood function from the weighted samples obtained from the E-step. To be more specific, Zhang, Rubin and Mealli (2009) and Frumento et al. (2012) proposed to use a multinomial logistic model for $\Pr(U_i|X_i, \theta_{W|X})$ and Normal linear models for $\Pr(Y_i(z)|U_i, X_i, \theta_{Y|U,X})$. If the parameter spaces of $\theta_{W|X}$ and $\theta_{Y|U,X}$ are distinct, then we can update them in the M-step separately: based on the weighted samples, $\theta_{W|X}^{[t+1]}$ can be obtained by fitting a weighted multinomial logistic regression, and $\theta_{Y|U,X}^{[t+1]}$ can be obtained by fitting weighted least squares with different combinations of $(Z_i = z, U_i = u)$.

In the analysis of randomized experiments with noncompliance, two assumptions are often invoked: monotonicity [i.e., $\Pr(U_i = d) = 0$], and exclusion restriction for noncompliers [i.e., $Y_i(1) = Y_i(0)$ for $U_i = a$ and n]. These assumptions ensure nonparametric identification of the CACE, and thus the result is not sensitive to parametric assumptions. Without these two

assumptions, subtle identification and inferential issues arise. Although the identifiability of parametric models is guaranteed by the identifiability of mixture models, the likelihood function may display pathological behavior (Ding et al., 2011, Frumento et al., 2016, Feller et al., 2016), and standard Frequentist inferential tools, for example, the bootstrap, do not apply. Without these two assumptions, the PCEs are only *partially* identified without additional parametric assumptions; the identification issue will be further discussed in the Bayesian perspective in Section 3.1.

2.3.3 *Truncation by death or missing data?* Though post-treatment variables do not have direct analogue in missing data analysis, principal stratification has several applications that shed new lights to some challenging missing data problems. One such application is the truncation by "death" problem. Consider a medical study that aims to evaluate the effect of a new treatment (Z_i) on the quality of life (Y_i); here, some patients may die before the outcome was measured and the post-treatment variable W_i is the survival status. Another example in labor economics is to evaluate the effects of a job training program (Z_i) on hourly wage (Y_i); here, some subjects may be unemployed and W_i is the employment status. A common feature of these two problems is that the outcome is not well defined or truncated if $W_i = 0$, that is, unit i is dead or unemployed. The name "truncation by death" comes from the original medical application (Rubin, 2006a).

Traditionally, the truncation by "death" problem was viewed as a standard missing data problem, with W_i being the missing data indicator. A famous model in this context is the Heckman selection model (Heckman, 1979), which consists of an outcome equation ($\mathbf{Y} \sim \mathbf{Z} + \mathbf{X}$) and a selection equation ($\mathbf{W} \sim \mathbf{Z} + \mathbf{X}$). This effectively assumes that the outcome is well defined for all units, which, however, is not easy to justify scientifically. Rubin (2006a) proposed to tackle this problem from a principal stratification perspective, namely, estimating the average causal effect among the units who would always survive under both treatment and control—the survivor average causal effect τ^{SACE}:

$$\tau^{\text{SACE}} = \mathbb{E}\{Y_i(1) - Y_i(0)|W_i(1) = W_i(0) = 1\}.$$

Estimating τ^{SACE} is again a missing data problem, but now the missingness arises from the latent stratum labels $\mathbf{U}$ rather than the original outcome $\mathbf{Y}$. This alternative view is arguably more scientifically relevant than

the traditional one. Zhang, Rubin and Mealli (2009) developed the EM-based estimation strategy in Section 2.3.2; recent extensions can be found in Ding et al. (2011), Frumento et al. (2012), Yang and Small (2016) and Ding and Lu (2017).

3. THE BAYESIAN PERSPECTIVE

The Bayesian paradigm offers a unified inferential framework for missing data and causal inference. Under the potential outcomes framework, each unit is associated with several quantities including the potential outcomes, some of which are observed and some are missing. Bayesian inference considers the observed values of these quantities to be realizations of random variables and the unobserved values to be unobserved random variables (Rubin, 1978), which are no different from unknown model parameters. Because inferences for the finite-sample and super-population estimands have subtle but important differences, below we will discuss them separately.

3.1 Finite-Sample Inference

Finite-sample causal estimands, defined as functions of the $N \times 2$ potential outcome matrix $\{\mathbf{Y}(0), \mathbf{Y}(1)\}$, can also be represented as functions of missing and observed potential outcomes, because $\tau^S = \tau(\mathbf{Y}(0), \mathbf{Y}(1)) = \tau(\mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}}, \mathbf{W})$. Finite-sample Bayesian causal inference centers around building a model to impute the missing potential outcomes given the observed data and then deriving the posterior distributions of the estimands.

Let $\Pr\{\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}, \mathbf{X}|\theta\}$ be the joint probability density function of the random variables for all units governed by a model parameter θ. For ease of exposition, we assume these random variables for each unit are independent and identically distributed conditional on θ, and thus we can factor the complete-data likelihood as $\prod_{i=1}^{N} \Pr\{Y_i(0), Y_i(1), W_i, X_i|\theta\}$. Imposing a prior distribution $\Pr(\theta)$ on the parameter, we obtain the joint posterior distribution of the missing data and the parameter as

$$
\begin{aligned}
&\Pr(\mathbf{Y}^{\text{mis}}, \theta|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}) \\
&\quad \propto \Pr(\theta) \prod_{i=1}^{N} \Pr(Y_i(0), Y_i(1), W_i, X_i|\theta).
\end{aligned}
\tag{12}
$$

Note that in equation (12) and the formulas below, we follow the convention of Bayesian statistics and use $\propto$ to denote "proportional to" a probability density,

dropping the normalizing constant. Formula (12)—followed immediately from the Bayes theorem—is fundamental to finite-sample Bayesian causal inference: One can first obtain the posterior distribution of $\mathbf{Y}^{\text{mis}}$ from (12), and then obtain the posterior distribution of $\tau^S = \tau(\mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}}, \mathbf{W})$ because $\mathbf{Y}^{\text{obs}}$ and $\mathbf{W}$ are known.

To make the role of (12) more explicit, we discuss two strategies to simulate the posterior distribution of $\mathbf{Y}^{\text{mis}}$, motivated by both computational and statistical considerations. The first strategy is based on data augmentation (Tanner and Wong, 1987) or more generally Gibbs sampling (Gelfand and Smith, 1990), that is, we iteratively simulate $\mathbf{Y}^{\text{mis}}$ and θ given each other and the observed data, based on $\Pr(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}, \theta)$ and $\Pr(\theta|\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X})$. Although this is a routine algorithm in Bayesian inference, explicating some steps is helpful for gaining insights. In particular, given the observed data and the parameter θ, the missing potential outcomes $\mathbf{Y}^{\text{mis}}$ have a posterior distribution as

$$
\begin{aligned}
&\Pr(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}, \theta) \\
&\quad \propto \Pr(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}, \mathbf{X}|\theta) \\
&\quad \propto \prod_{i=1}^{N} \Pr(W_i|Y_i(0), Y_i(1), X_i, \theta) \\
&\qquad \cdot \Pr(Y_i(0), Y_i(1)|X_i, \theta) \\
&\qquad \cdot \Pr(X_i|\theta).
\end{aligned} \tag{13}
$$

It is common to parameterize the assignment mechanism, the distribution of potential outcomes and the distribution of covariates with different and a priori independent sets of parameters, say $\theta_{W|X}$, $\theta_{Y|X}$ and θ_X, respectively. Then, given unconfoundedness, the assignment mechanism $\Pr(W_i|Y_i(0), Y_i(1), \mathbf{X}_i, \theta)$ and the covariate distribution $\Pr(X_i|\theta)$ drop out in (13), which simplifies to

$$
\begin{aligned}
&\Pr(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}, \theta) \\
&\quad \propto \prod_{i:W_i=1} \Pr(Y_i(0), Y_i(1)|X_i, \theta_{Y|X}) \\
&\qquad \cdot \prod_{i:W_i=0} \Pr(Y_i(0), Y_i(1)|X_i, \theta_{Y|X}) \\
&\quad \propto \prod_{i:W_i=1} \Pr(Y_i(0)|Y_i(1), X_i, \theta_{Y|X}) \\
&\qquad \cdot \prod_{i:W_i=0} \Pr(Y_i(1)|Y_i(0), X_i, \theta_{Y|X}).
\end{aligned} \tag{14}
$$

Because the posterior distribution (14) factors into N terms, for treated units we impute the missing control

potential outcomes from $\Pr(Y_i(0)|Y_i(1), X_i, \theta_{Y|X})$, and for control units we impute the missing treatment potential outcomes from $\Pr(Y_i(1)|Y_i(0), X_i, \theta_{Y|X})$. Imputing the missing potential outcomes $\mathbf{Y}^{\text{mis}}$ depends crucially on the joint distribution of $\{Y_i(1), Y_i(0)\}$ given X_i. Given the observed data and the imputed $\mathbf{Y}^{\text{mis}}$, the posterior distribution of $\theta_{Y|X}$ can be obtained by a complete-data analysis based on $\Pr(\theta_{Y|X}|\mathbf{Y}(1), \mathbf{Y}(0), \mathbf{X}) \propto \Pr(\theta_{Y|X}) \prod_{i=1}^{N} \Pr(Y_i(1), Y_i(0)|X_i, \theta_{Y|X})$.

The above general framework was first proposed by Rubin (1978) and has been widely adopted in the literature; recently Heckman, Lopes and Piatek (2014) extended it to more specific econometric models. However, this strategy has a limitation of mixing fully identifiable and non (or weakly) identifiable parameters. Due to the simultaneous presentation of several inferential frameworks in this article, we need to clarify the different notions of identifiability in each framework. Under the Frequentist paradigm, a parameter is identifiable if it can be expressed as a function of the distribution of the observed data (Bickel and Doksum, 2015). In other words, a parameter is identifiable in the Frequentist sense if two distinct values of it give two different distributions of the observed data. Under the Bayesian paradigm, there is no consensus. For example, Lindley (1972) argued that in Bayesian analysis, all parameters are identifiable because with proper prior distributions, posterior distributions are always proper. However, Gustafson (2015) argued that a parameter is only *weakly* or *partially* identifiable, if a substantial region of its posterior distribution is flat, or its posterior distribution depends crucially on its prior distribution even with large samples. Specifically, because $Y_i(1)$ and $Y_i(0)$ are never jointly observed, the data provide little information about the parameter that governs the association between $Y_i(1)$ and $Y_i(0)$. Consequently, the posterior distribution of τ^S will be sensitive to its prior distribution. Therefore, it may be more sensible to isolate the parameter—denoted by θ^{m}—that governs the marginal distributions from the parameter—denoted by θ^{a}—that governs the association between $Y_i(1)$ and $Y_i(0)$. This motivates the following strategy to simulate the posterior distribution of $\mathbf{Y}^{\text{mis}}$.

The second strategy is based on the definition of conditional probability, that is,

$$
\begin{aligned}
&\Pr(\mathbf{Y}^{\text{mis}}, \theta|\mathbf{O}^{\text{obs}}) \\
&\quad = \Pr(\theta|\mathbf{O}^{\text{obs}}) \Pr(\mathbf{Y}^{\text{mis}}|\theta, \mathbf{O}^{\text{obs}}),
\end{aligned}
$$

where $\mathbf{O}^{\text{obs}} = (\mathbf{X}, \mathbf{Y}^{\text{obs}}, \mathbf{W})$. Here, we first simulate θ given the observed data, $\Pr(\theta|\mathbf{O}^{\text{obs}})$, and then simulate

$\mathbf{Y}^{\mathrm{mis}}$ given θ and the observed data, $\Pr(\mathbf{Y}^{\mathrm{mis}}|\theta, \mathbf{O}^{\mathrm{obs}})$. Following the above arguments, we partition $\theta_{Y|X}$ into $\theta_{Y|X}^{\mathrm{m}}$ and $\theta_{Y|X}^{\mathrm{a}}$, and impose independent priors on them. Therefore, the posterior distribution of θ becomes

$$\Pr(\theta|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W}, \mathbf{X})$$

$$(15) \quad \propto \Pr(\theta_{Y|X}^{\mathrm{a}})\Pr(\theta_{Y|X}^{\mathrm{m}}) \prod_{W_i=1} \Pr(Y_i(1)|X_i, \theta_{Y|X}^{\mathrm{m}})$$

$$\cdot \prod_{W_i=0} \Pr(Y_i(0)|X_i, \theta_{Y|X}^{\mathrm{m}}).$$

Not surprisingly, the posterior distribution of $\theta_{Y|X}^{\mathrm{m}}$ is updated by the likelihood, but the posterior of $\theta_{Y|X}^{\mathrm{a}}$ remains the same as its prior. This is due to the a priori independence between $\theta_{Y|X}^{\mathrm{m}}$ and $\theta_{Y|X}^{\mathrm{a}}$; otherwise the posterior of $\theta_{Y|X}^{\mathrm{a}}$ will be updated *indirectly* via its dependence on $\theta_{Y|X}^{\mathrm{m}}$, as pointed out by Gustafson (2009). Richardson, Evans and Robins (2010) suggested transparent parametrization by separating identifiable and nonidentifiable parameters in Bayesian causal inference; Ding and Dasgupta (2016) suggested sensitivity analysis by varying $\theta_{Y|X}^{\mathrm{a}}$ in a certain range (more details in Example 2 below). After drawing $\theta_{Y|X}$ from its posterior distribution (15), we can impute the missing potential outcomes $\mathbf{Y}^{\mathrm{mis}}$ using the same formula as (14).

To illustrate the idea, we discuss two examples arising from the Bayesian analysis of completely randomized experiments.

EXAMPLE 1 (Two-by-two table). In a completely randomized experiment with treatment W and binary outcome Y, the joint potential outcomes $\{Y_i(1), Y_i(0)\}$ can take four values with counts $N_{y_1 y_0} = \#\{i : Y_i(1) = y_1, Y_i(0) = y_0\}$ for $y_1, y_0 = 0, 1$, and the observed data can be summarized by a two-by-two table with counts $n_{wy} = \#\{i : W_i = w, Y_i^{\mathrm{obs}} = y\}$ for $w, y = 0, 1$.

Following the first strategy, we can impose a multinomial model for $\{Y_i(1), Y_i(0)\}$ with a Dirichlet prior for the probabilities

$$\{Y_i(1), Y_i(0)\}|(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$$

$$\sim \mathrm{Multinomial}(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}),$$

$$(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$$

$$\sim \mathrm{Dirichlet}(\alpha_{11}, \alpha_{10}, \alpha_{01}, \alpha_{00}).$$

Given the observed data and the parameter, it is straightforward to impute the missing potential outcomes. For example, for units with $W_i = 1$ and $Y_i^{\mathrm{obs}} = 1$, we draw the missing $Y_i(0)$ from Bernoulli$(\pi_{11}/(\pi_{11} + \pi_{10}))$; for other three types of units, we can

similarly draw their missing potential outcomes. After imputing the missing potential outcomes, the counts $(N_{11}, N_{10}, N_{01}, N_{00})$ are known, and we can then draw the parameters from their posterior:

$$(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})|(N_{11}, N_{10}, N_{01}, N_{00})$$

$$\sim \mathrm{Dirichlet}(\alpha_{11} + N_{11}, \alpha_{10} + N_{10},$$

$$\alpha_{01} + N_{01}, \alpha_{00} + N_{00}).$$

This data augmentation scheme gives us the posterior distributions of $(N_{11}, N_{10}, N_{01}, N_{00})$, which immediately imply the posterior distribution of $\tau^S = (N_{10} - N_{01})/N$. However, the posterior distribution of τ^S is sensitive to the choice of the hyperparameters $(\alpha_{11}, \alpha_{10}, \alpha_{01}, \alpha_{00})$, because they determine the a priori dependence between the two potential outcomes; see Ding (2014), Section 3, for the consequence of using the Jeffreys "noninformative" prior: $(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}) \sim \mathrm{Dirichlet}(1/2, 1/2, 1/2, 1/2)$.

Following the second strategy, Ding and Dasgupta (2016) reparametrize the joint distribution of the potential outcomes as $(\pi_{1+}, \pi_{+1}, \gamma)$, where $\pi_{1+} = \pi_{11} + \pi_{10} = \Pr(Y_i(1) = 1)$ and $\pi_{+1} = \pi_{11} + \pi_{01} = \Pr(Y_i(0) = 1)$ govern the marginal distributions, and $\gamma = \Pr(Y_i(1) = 1|Y_i(0) = 1)/\Pr(Y_i(1) = 1|Y_i(0) = 0)$ governs the association. From $(\pi_{1+}, \pi_{+1}, \gamma)$, we can uniquely determine $(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$. Ding and Dasgupta (2016) advocate treating γ as the sensitivity parameter and obtaining posterior distribution of τ^S for a fixed γ. If we postulate independent priors $\pi_{1+} \sim \mathrm{Beta}(\alpha_1, \beta_1)$ and $\pi_{+1} \sim \mathrm{Beta}(\alpha_0, \beta_0)$, the posterior distributions are $\pi_{1+} \sim \mathrm{Beta}(\alpha_1 + n_{11}, \beta_1 + n_{10})$ and $\pi_{+1} \sim \mathrm{Beta}(\alpha_0 + n_{01}, \beta_0 + n_{00})$. After drawing π_{1+} and π_{+1}, we can impute all the missing potential outcomes. For example, for units with $W_i = 1$ and $Y_i^{\mathrm{obs}} = 1$, we draw $Y_i(0)$ from a Bernoulli$(\gamma \pi_{+1}/(1 - \pi_{+1} + \gamma \pi_{+1}))$; for other three types of units, we can similarly draw their missing potential outcomes. Therefore, we can obtain the posterior distribution of τ^S for a fixed γ. Varying γ in a selected range yields a Bayesian sensitivity analysis.

EXAMPLE 2 (Bayesian post-stratification). For illustration purposes, we present an example of model-based covariate adjustment in a completely randomized experiment, which is the Bayesian analogue of the Frequentist post-stratification (Miratrix, Sekhon and Yu, 2013). For each unit i, we observed the binary treatment indicator W_i, outcome Y_i and a discrete covariate X_i taking values in $\{1, \ldots, K\}$. Assume the following joint model for the potential outcomes given the covari-

ate:

$$\binom{Y_i(1)}{Y_i(0)}\Big|(X_i = k, \theta_{Y|X})$$

$$\sim N\left(\binom{\mu_{1[k]}}{\mu_{0[k]}}, \begin{pmatrix} \sigma_{1[k]}^2 & \rho_{[k]}\sigma_{1[k]}\sigma_{0[k]} \\ \rho_{[k]}\sigma_{1[k]}\sigma_{0[k]} & \sigma_{0[k]}^2 \end{pmatrix}\right),$$

$$(k = 1, \ldots, K; i = 1, \ldots, N),$$

where $\theta_{Y|X}^{\mathrm{m}} = \{\mu_{1[k]}, \mu_{0[k]}, \sigma_{1[k]}^2, \sigma_{0[k]}^2\}_{k=1}^K$ and $\theta_{Y|X}^{\mathrm{a}} = \{\rho_{[k]}\}_{k=1}^K$. The observed likelihood factors into two parts: the data in treatment group $\{(X_i, Y_i^{\mathrm{obs}}) : W_i = 1\}$ contribute to the likelihood of $\{\mu_{1[k]}, \sigma_{1[k]}^2\}_{k=1}^K$, and the data in the control group $\{(X_i, Y_i^{\mathrm{obs}}) : W_i = 0\}$ contribute to the likelihood of $\{\mu_{0[k]}, \sigma_{0[k]}^2\}_{k=1}^K$. Importantly, the observed likelihood does not depend on $\theta_{Y|X}^{\mathrm{a}}$.

Imposing independent priors on all the parameters, we can easily obtain their posterior distributions. For example, the conventional noninformative prior for the Normal models yield conjugate posteriors, because within each stratum $X_i = k$ it is a standard Normal model. For convenience of description, we define $n_{w[k]}$ as the number of units, and $(\bar{Y}_{w[k]}^{\mathrm{obs}}, s_{w[k]}^2)$ as the sample mean and variance of the outcomes under treatment w within stratum $X_i = k$. If $p(\mu_{w[k]}) \propto 1$ and $p(\sigma_{w[k]}^2) \propto 1/\sigma_{w[k]}^2$ for $w = 0, 1$ and $k = 1, \ldots, K$, then according to Gelman et al. [2014, Section 3.2], the posterior distributions of the parameters are

$$\sigma_{w[k]}^2|\mathbf{O}^{\mathrm{obs}} \sim (n_{w[k]} - 1)s_{w[k]}^2/\chi_{n_{w[k]}-1}^2,$$
(16)
$$\mu_{w[k]}|\sigma_{w[k]}^2, \mathbf{O}^{\mathrm{obs}} \sim N(\bar{Y}_{w[k]}^{\mathrm{obs}}, \sigma_{w[k]}^2/n_{w[k]}).$$

For fixed values of $\theta_{Y|X}^{\mathrm{a}}$ and given each draw of $\theta_{Y|X}^{\mathrm{m}}$, we can impute the missing potential outcomes as follows: for treated units ($W_i = 1$) within stratum $X_i = k$, we draw

$$Y_i(0)|X_i = k, W_i = 1, Y_i^{\mathrm{obs}}, \theta_{Y|X}$$

$$\sim N\Big(\mu_{0[k]}$$

$$+ \rho_{[k]}\frac{\sigma_{0[k]}}{\sigma_{1[k]}}(Y_i^{\mathrm{obs}} - \mu_{1[k]}), \sigma_{0[k]}^2(1 - \rho_{[k]}^2)\Big),$$

and for control units ($W_i = 0$) within stratum $X_i = k$, we draw

$$Y_i(1)|X_i = k, W_i = 1, Y_i^{\mathrm{obs}}, \theta_{Y|X}$$

$$\sim N\Big(\mu_{1[k]}$$

$$+ \rho_{[k]}\frac{\sigma_{1[k]}}{\sigma_{0[k]}}(Y_i^{\mathrm{obs}} - \mu_{0[k]}), \sigma_{1[k]}^2(1 - \rho_{[k]}^2)\Big).$$

Consequently, we obtain the posterior distribution of τ^S. In practice, we suggest varying the $\rho_{[k]}$'s from 0 to 1, which correspond to conditionally independent potential outcomes and perfectly correlated potential outcomes.

Example 2 deserves some further discussions. First, for simplicity we can further assume that $\rho_{[k]} = \rho$ and, therefore, reduce the number of sensitivity parameters to one. Second, the above strategy works well for small K and large $n_{w[k]}$'s. In particular, the posterior distributions (16) are well defined if $n_{w[k]} \geq 1$, and the marginal posterior mean of $\mu_{w[k]}$ is finite is $n_{w[k]} \geq 3$. For large K and small $n_{w[k]}$'s, we have a large number of parameters and, therefore, need to impose a more sophisticated prior on $\theta_{Y|X}^{\mathrm{m}}$. Ideally, we want to choose a prior that yields good Frequentist properties under the randomization inference framework, because the original design is a completely randomized experiment. This deserves further research. Third, with a discrete covariate, we can also define the ATE over sub-populations:

$$\tau_k^S = \frac{1}{N_{[k]}} \sum_{i:X_i=k} \{Y_i(1) - Y_i(0)\},$$

where $N_{[k]} = n_{1[k]} + n_{0[k]}$ is the number of units within covariate stratum $X_i = k$. According to Example 2, obtaining the posterior distribution of τ_k^S is straightforward.

3.2 Super-Population Inference

In general, a super-population causal parameter is a function of the model parameters for the joint distribution of $\{X_i, Y_i(1), Y_i(0)\}$, written as $\tau = \tau(\theta_{Y|X}, \theta_X)$. Therefore, Bayesian inference for super-population causal parameters reduces to obtaining posterior distributions of $\theta_{Y|X}$ and θ_X. Examples of such causal parameters include the subgroup treatment effect

$$\tau^P(x) \equiv \mathbb{E}\{Y(1)|X = x; \theta_{Y|X}^{\mathrm{m}}\}$$
(17)
$$- \mathbb{E}\{Y(0)|X = x; \theta_{Y|X}^{\mathrm{m}}\},$$

the PATE

$$\tau^P \equiv \mathbb{E}\{Y_i(1) - Y_i(0)\}$$
(18)
$$= \int \tau^P(x; \theta_{Y|X}^{\mathrm{m}})F_X(\mathrm{d}x; \theta_X).$$

In most cases, however, we are unwilling to model the possibly multidimensional pretreatment covariate X_i and, therefore, condition on the observed values of the covariates, which is equivalent to replace θ_X with $\widehat{\mathbb{F}}_X$,

the empirical distribution of $\{X_i\}_{i=1}^N$. Therefore, most Bayesian causal inferences in fact focus on the conditional ATE:

$$
(19) \quad
\begin{aligned}
\tau^{\mathbf{X}} &\equiv \int \tau^P\big(x; \theta_{Y|X}^{\mathrm{m}}\big) \widehat{\mathbb{F}}_X(\mathrm{d}x) \\
&= N^{-1} \sum_{i=1}^N \tau^P\big(X_i; \theta_{Y|X}^{\mathrm{m}}\big).
\end{aligned}
$$

Note that in general $\tau^{\mathbf{X}}$ is neither the PATE nor the SATE. Because we are often unwilling to model X, we omit the discussion of τ^P and focus on $\tau^{\mathbf{X}}$ in this subsection.

It is worth giving two concrete examples for the causal estimands above. Here, we assume that the covariate vector X_i contains an intercept. For a continuous outcome, if $Y_i(w)|X_i \sim N(\beta_w' X_i, \sigma_w^2)$ for $w = 1, 0$, then $\tau^P(x) = (\beta_1 - \beta_0)'x$ and $\tau^{\mathbf{X}} = N^{-1} \cdot \sum_{i=1}^N (\beta_1 - \beta_0)' X_i = (\beta_1 - \beta_0)' \bar{X}$. For a binary outcome, if $\Pr(Y_i(w) = 1|X_i) = F(\beta_w' X_i)$ for $w = 1, 0$ with $F(\cdot)$ being the link function [e.g., the standard Normal distribution $F(\cdot) = \Phi(\cdot)$ gives the probit model], then $\tau^P(x) = F(\beta_1' x) - F(\beta_0' x)$ and $\tau^{\mathbf{X}} = N^{-1} \sum_{i=1}^N \{F(\beta_1' X_i) - F(\beta_0' X_i)\}$.

A salient feature of the estimands in (17)–(19) is that they depend only on $\theta_{Y|X}^{\mathrm{m}}$ but not on $\theta_{Y|X}^{\mathrm{a}}$. For this type of estimands, we do not need to impose a joint model of $\{Y_i(1), Y_i(0)\}$ given X_i, because the likelihood, and thus the posterior distribution of $\theta_{Y|X}^{\mathrm{m}}$ requires only the specification of the marginal distributions of $Y_i(w)$ given X_i for $w = 0, 1$. Bayesian inference for estimands (17)–(19) is straightforward once we obtain the posterior distributions of the parameter $\theta_{Y|X}^{\mathrm{m}}$ following the second strategy in Section 3.1.

Moreover, we can infer parameters that depend on the joint distribution of the potential outcomes. For example, we can consider $\delta_1 = \Pr(Y_i(1) \geq Y_i(0))$ and its conditional version $\delta_1^{\mathbf{X}} = N^{-1} \sum_{i=1}^N \delta_1(X_i)$, where

$$
\delta_1(x) = \Pr\big(Y_i(1) > Y_i(0) | X_i = x, \theta_{Y|X}^{\mathrm{m}}, \theta_{Y|X}^{\mathrm{a}}\big).
$$

We give an example to illustrate inferring $\delta_1^{\mathbf{X}}$ under a Normal linear model as in Heckman, Lopes and Piatek (2014).

EXAMPLE 3 (Normal linear model). Assume

$$
(20) \quad
\begin{pmatrix} Y_i(1) \\ Y_i(0) \end{pmatrix} \Big| (X_i, \theta_{Y|X}) \\
\sim N\left(\begin{pmatrix} \beta_1' X_i \\ \beta_0' X_i \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_0 \\ \rho\sigma_1\sigma_0 & \sigma_0^2 \end{pmatrix} \right), \\
(i = 1, \ldots, N),
$$

where $\theta_{Y|X}^{\mathrm{m}} = (\beta_1, \beta_0, \sigma_1^2, \sigma_0^2)$ and $\theta_{Y|X}^{\mathrm{a}} = \rho$. The model (20) implies

$$
(21) \quad Y_i(w)|X_i, \theta_{Y|X}^{\mathrm{m}} \sim N\big(\beta_w' X_i, \sigma_w^2\big), \quad (w = 0, 1)
$$

and

$$
(22) \quad
\begin{aligned}
&Y_i(1) - Y_i(0)|X_i, \theta_{Y|X} \\
&\sim N\big((\beta_1 - \beta_0)' X_i, \sigma_1^2 + \sigma_0^2 - 2\rho\sigma_1\sigma_0\big).
\end{aligned}
$$

Therefore, we can derive from (22) that

$$
\delta_1^{\mathbf{X}} = \frac{1}{N} \sum_{i=1}^N \Phi\left\{ \frac{(\beta_1 - \beta_0)' X_i}{(\sigma_1^2 + \sigma_0^2 - 2\rho\sigma_1\sigma_0)^{1/2}} \right\}.
$$

Based on the marginal model (21) we can obtain the posterior distribution of $\theta_{Y|X}^{\mathrm{m}}$, for example, via standard Normal posterior calculations. Since the observed data do not contain any information about ρ, we can vary it from 0 to 1 and obtain the posterior distribution of $\delta_1^{\mathbf{X}}$ for each fixed ρ.

The Bayesian framework offers a unified and flexible approach to inferring causal parameters in complex settings. Indeed, there is a rapidly growing literature in applying advanced Bayesian models and methods—such as Bayesian nonparametric methods, Bayesian model selection and model averaging—to causal inference. A comprehensive review of the Bayesian approach to causal inference is highly relevant, but is beyond the scope of this paper.

4. THE FISHERIAN RANDOMIZATION PERSPECTIVE

The Fisherian randomization perspective to causal inference focuses on p-values under null hypotheses obtained by comparing the observed values of test statistics with their randomization distribution (Fisher, 1935). Broadly speaking, the Fisherian perspective belongs to the Frequentist perspective because the p-value is commonly viewed as a Frequentist notion. For its special emphases and historical reasons, we classify the Fisherian perspective as a separate inferential framework. This perspective is rather unique to causal inference because standard missing data problems do not involve comparisons of different treatment groups. The Fisherian framework is also closely related to the Bayesian perspective in that a key step is to impute all the missing potential outcomes.

Because the Fisherian framework has been largely developed in randomized studies, we will focus on the finite-sample inference.

4.1 Fisher Randomization Test of the Sharp Null Hypothesis

The Fisherian randomization test considers a finite population of N units with N_1 receiving treatment and N_0 receiving control. For the ease of exposition, we discuss the case with a known propensity score $\Pr(\mathbf{W}|\mathbf{X})$ such as in (stratified) completely randomized experiments. We first present a general form of the Fisher randomization test for a sharp null hypothesis that $Y_i(1) - Y_i(0) = \tau_i$ with a known constant vector $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)$. The null hypothesis is sharp because it allows for imputing all the missing potential outcomes: for a unit with $W_i = 1$, we have $Y_i(1) = Y_i^{\mathrm{obs}}$ and $Y_i(0) = Y_i^{\mathrm{mis}} = Y_i^{\mathrm{obs}} - \tau_i$; for a unit with $W_i = 0$, we have $Y_i(0) = Y_i^{\mathrm{obs}}$ and $Y_i(1) = Y_i^{\mathrm{mis}} = Y_i^{\mathrm{obs}} + \tau_i$. Originally, Fisher (1935) considered the sharp null $H_0^{\#}$ with $\tau_i = 0$ for all $i = 1, \ldots, N$, or equivalently $\mathbf{Y}(1) = \mathbf{Y}(0) = \mathbf{Y}^{\mathrm{obs}}$.

In general, a test statistic $t(\mathbf{W}, \mathbf{Y}(1), \mathbf{Y}(0), \mathbf{X})$ is a function of the treatment assignment vector, potential outcomes and pretreatment covariates. For example, it can be the difference-in-means of the outcomes, $N_1^{-1} \sum_{i=1}^N W_i Y_i(1) - N_0^{-1} \sum_{i=1}^N (1 - W_i) Y_i(0)$, or the difference-in-means of the residuals from linear regressions of $\mathbf{Y}(1)$ and $\mathbf{Y}(0)$ on $\mathbf{X}$, $N_1^{-1} \sum_{i=1}^N W_i \varepsilon_{i1} - N_0^{-1} \sum_{i=1}^N (1 - W_i) \varepsilon_{i0}$, where ε_1s and ε_0s consist of the residuals from the linear regressions of $\mathbf{Y}(1)$ and $\mathbf{Y}(0)$ on $\mathbf{X}$ without an intercept, respectively (Tukey, 1993, Rosenbaum, 2002a). Because $\mathbf{X}$ is fixed and $\{\mathbf{Y}(1), \mathbf{Y}(0)\}$ are known under the sharp null, the distribution of $t(\mathbf{W}, \mathbf{Y}(1), \mathbf{Y}(0), \mathbf{X})$ is determined by the only random component $\mathbf{W}$. Therefore, we can calculate or simulate the randomization distribution of the test statistic according to $\Pr(\mathbf{W}|\mathbf{X})$, and then obtain the p-value $\Pr(\tilde{t} \geq t)$, defined as the tail probability of the test statistic t with respect to its randomization distribution $\tilde{t}$.

4.2 Fisherian p-Value as a Posterior Predictive p-Value

We first review the posterior predictive p-values. Although general forms exist (Gelman, Meng and Stern, 1996), we use Meng's (1994) formulation of the posterior predictive p-value (ppp) and tailor it to the finite-sample causal inference. Recall that $\mathbf{O}^{\mathrm{obs}}$ and $\mathbf{O}^{\mathrm{mis}}$ denote the observed and missing data, respectively. Consider testing a null hypothesis H_0.

If $\mathbf{O}^{\mathrm{mis}}$ were known, then we can choose a test statistic $t(\mathbf{O}^{\mathrm{obs}})$ to measure a deviation from H_0, and obtain the p-value $\mathrm{p}(\mathbf{O}^{\mathrm{mis}})$. We highlight the dependence

of the p-value on $\mathbf{O}^{\mathrm{mis}}$, because the distribution of $t(\mathbf{O}^{\mathrm{obs}})$ depends on $\mathbf{O}^{\mathrm{mis}}$ in general. The ppp-value with missing potential outcomes is defined as the mean of $\mathrm{p}(\mathbf{O}^{\mathrm{mis}})$ over the posterior distribution of $\mathbf{O}^{\mathrm{mis}}$ given the observed data $\mathbf{O}^{\mathrm{obs}}$, $\Pi(\mathbf{O}^{\mathrm{obs}}|\mathbf{O}^{\mathrm{obs}})$,

$$(23) \qquad \mathrm{ppp} = \int \mathrm{p}(\mathbf{O}^{\mathrm{mis}}) \Pi(\mathbf{O}^{\mathrm{mis}}|\mathrm{d}\mathbf{O}^{\mathrm{obs}}).$$

In order to obtain the posterior distribution of $\mathbf{O}^{\mathrm{mis}}$, we often need to invoke a model for the complete data $\Pr(\mathbf{O}^{\mathrm{obs}}, \mathbf{O}^{\mathrm{mis}}|\theta)$ and a prior $\Pr(\theta)$. The joint posterior of $(\mathbf{O}^{\mathrm{mis}}, \theta)$ can be obtained from

$$\Pr(\mathbf{O}^{\mathrm{mis}}, \theta|\mathbf{O}^{\mathrm{obs}}) \propto \Pr(\theta) \Pr(\mathbf{O}^{\mathrm{obs}}, \mathbf{O}^{\mathrm{mis}}|\theta),$$

and the posterior $\Pi(\mathbf{O}^{\mathrm{mis}}|\mathbf{O}^{\mathrm{obs}})$ can be obtained by marginalizing over θ.

The above formulation is generic and applicable to general finite-sample causal inference. Below we discuss an application, using the Fisher randomization test to obtain $\mathrm{p}(\mathbf{O}^{\mathrm{mis}})$ and using a Bayesian model to obtain the posterior distribution of the missing potential outcomes. Therefore, the ppp discussed in this subsection has both Bayesian and Fisherian favors.

The pure randomization-based test was often viewed as a model-free robust procedure that had no connection to the Bayesian inference. However, Rubin (1984) first interpreted the Fisher randomization test as a Bayesian posterior predictive check. We assume that $\Pr(\mathbf{W}|\mathbf{X})$ has a known probabilistic law as in randomized experiments. With a known unit-level treatment effect vector $\boldsymbol{\tau}$, the posterior distribution $\Pi(\mathbf{O}^{\mathrm{mis}}|\mathbf{O}^{\mathrm{obs}})$ contains point masses at the imputed values of the missing potential outcomes. Therefore, the posterior predictive replicates can be obtained by drawing $\mathbf{W}$ according to its probabilistic law without changing the values of $\{\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)\}$, and the ppp using statistic $t(\mathbf{W}, \mathbf{Y}(1), \mathbf{Y}(0), \mathbf{X})$ is identical to that obtained from the Fisher randomization test in Section 4.1.

4.3 Nonsharp Null Hypotheses

Although the Bayesian interpretation of the Fisher randomization test in Section 4.2 is "intellectually pleasing" (Rubin, 2005), its real advantages are not obvious in the case of a sharp null hypothesis. When the null hypothesis is not sharp, it is often challenging to conduct exact randomization tests without sacrificing statistical power (Nolen and Hudgens, 2011, Ding, Feller and Miratrix, 2016). Fortunately, we can still obtain ppp's in these cases, because we can first obtain p-values for fixed values of the unknown missing potential outcomes, and then average the p-values over

the posterior distributions of these unknown potential outcomes. We give two examples to illustrate this general approach.

EXAMPLE 4 (Testing treatment effect variation). We are interested in testing whether the unit-level treatment effects are constant (Ding, Feller and Miratrix, 2016), a model that is often assumed in randomization-based causal inference (Rosenbaum, 2002b):

$$H_0^C : Y_i(1) - Y_i(0) = \tau$$

$$\text{for some } \tau \text{ and for all } i = 1, \dots, N.$$

Because the null hypothesis H_0^C depends on an unknown parameter τ, it is not sharp, and thus the missing potential outcomes cannot be imputed as in Section 4.1. However, for a fixed value of τ, the null hypothesis that $Y_i(1) - Y_i(0) = \tau$ for all unit i is sharp, and all the missing potential outcomes can be imputed: for a unit with $W_i = 1$, we have $Y_i(1) = Y_i^{\text{obs}}$ and $Y_i(0) = Y_i^{\text{mis}} = Y_i^{\text{obs}} - \tau$; for a unit with $W_i = 0$, we have $Y_i(0) = Y_i^{\text{obs}}$ and $Y_i(1) = Y_i^{\text{mis}} = Y_i^{\text{obs}} + \tau$. Ding, Feller and Miratrix (2016) suggested using the following shifted Kolmogorov–Smirnov statistic for H_0^C:

$$
\begin{aligned}
& t_{\text{SKS}}(\mathbf{W}, \mathbf{Y}(1), \mathbf{Y}(0), \mathbf{X}) \\
& = \max_y \left| N_1^{-1} \sum_{W_i=1} I(Y_i^{\text{obs}} - \widehat{\tau} \le y) \right. \\
& \left. \quad - N_0^{-1} \sum_{W_i=0} I(Y_i^{\text{obs}} \le y) \right|,
\end{aligned}
$$

where $\widehat{\tau}$ is the difference in means of the outcomes in treatment and control groups. For a fixed value of τ, we can first impute all the missing potential outcomes, and then obtain the p-value $\text{p}(\tau) = \Pr(\widetilde{t}_{\text{SKS}} \ge t_{\text{SKS}} | \tau)$ using the Fisher randomization test, where $\widetilde{t}_{\text{SKS}}$ represents the randomization distribution of the test statistic t_{SKS} that can be simulated by Monte Carlo. Note that in this example, because under H_0^C there is a one-to-one map between τ and $\mathbf{O}^{\text{mis}} = \mathbf{Y}^{\text{mis}}$, we use the simple notation $\text{p}(\tau)$ for $\text{p}(\mathbf{O}^{\text{mis}})$.

However, we do not know τ and need to obtain its posterior distribution. Assume that the first part of the joint model $\Pr(\mathbf{W}|\mathbf{X})$ is known as in (stratified) completely randomized experiments. We need only to model the second part $\Pr(\mathbf{Y}(1), \mathbf{Y}(0)|\mathbf{X}, \tau, \theta) = \Pr(\mathbf{Y}(0)|\theta) \times I\{\mathbf{Y}(1) = \mathbf{Y}(0) + \tau \mathbf{1}_N\}$ and impose priors on (τ, θ). For notational simplicity, we illustrate the idea with a Normal linear model $\mathbf{Y}(0) = \mathbf{X}\beta + \varepsilon$ with an intercept, where $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$ and $\theta = (\beta, \sigma^2)$. As a result, the observed outcomes follow

$\mathbf{Y}^{\text{obs}} = \tau \mathbf{W} + \mathbf{X}\beta + \varepsilon$, and under the usual Normal and inverse-χ^2 priors for (τ, β, σ^2), the posterior of τ can be easily obtained in closed form (Gelman et al., 2014, Chapter 14). The final ppp-value is obtained by averaging $\text{p}(\tau)$ over the posterior distribution $\Pi(\tau|\mathbf{O}^{\text{obs}})$, which is $\text{ppp}(H_0^C) = \int p(\tau) \Pi(\text{d}\tau|\mathbf{O}^{\text{obs}})$.

Ding, Feller and Miratrix (2016) also discussed several alternative test statistics. In practice, we may also want to impose more flexible outcome models beyond the Normal linear model, but the essence of the ppp remains the same.

EXAMPLE 5 (Testing treatment effects among compliers). We revisit the principal stratification approach to noncompliance (Section 2.3) under the Fisherian perspective. Recall that for unit i, let X_i, Z_i, $W_i(1)$, $W_i(0)$, $Y_i(1)$ and $Y_i(0)$ be covariate, treatment assignment, potential values of the actual treatment and the outcome, respectively. Recall that $U_i = \{W_i(1), W_i(0)\}$ is the latent principal stratum. Rubin (1998) and Rosenbaum (2002b) considered testing the following null hypothesis for the compliers:

$$H_0^c : Y_i(1) = Y_i(0) \quad (\text{for } U_i = c, i = 1, \dots, N).$$

Assuming monotonicity and exclusion restriction, this null hypothesis is equivalent to Fisher's sharp null hypothesis $H_0^\#$ in Section 4.1 that the treatment does not affect the outcomes of any units. This equivalence implies that we can simply conduct the usual Fisher randomization test discussed in Section 4.1 for H_0^c based on the data $(\mathbf{X}, \mathbf{Z}, \mathbf{Y}^{\text{obs}})$, ignoring the information of noncompliance $\mathbf{W}^{\text{obs}}$. However, this exact randomization test does not make full use of the observed data, and consequently can lose power in some cases.

Define $(\bar{Y}_1^{\text{obs}}, \bar{Y}_0^{\text{obs}})$ and $(\bar{W}_1^{\text{obs}}, \bar{W}_0^{\text{obs}})$ as the observed means of the treatments received and outcomes under treatment and control assignments. Two commonly used estimators for the CACE are (i) the Wald estimator $t_{\text{Wald}} = (\bar{Y}_1 - \bar{Y}_0)/(\bar{W}_1 - \bar{W}_0)$ (Angrist, Imbens and Rubin, 1996), and (ii) the 2SLS estimator t_{TSLS} that is the coefficient of $\mathbf{W}^{\text{obs}}$ in the regression $\mathbf{Y}^{\text{obs}} \sim \mathbf{W}^{\text{obs}} + \mathbf{X}$ using $\mathbf{Z}$ as an instrument for $\mathbf{W}^{\text{obs}}$ (Angrist and Pischke, 2008). It is intuitive to use one of them as a test statistic in the randomization test. Unfortunately, the null distributions of these test statistics depend on the unknown potential values of the treatment received $\mathbf{W}^{\text{mis}}$: for unit with $Z_i = 1$, we have $W_i^{\text{mis}} = W_i(0)$, and for units with $Z_i = 0$, we have $W_i^{\text{mis}} = W_i(1)$. If $\mathbf{W}^{\text{mis}}$ were known, then we can simulate the distribution of the test statistic t (e.g., t_{Wald} or t_{TSLS}) to obtain the p-value $\text{p}(\mathbf{W}^{\text{mis}}) = \Pr(\widetilde{t} \ge t | \mathbf{W}^{\text{mis}})$, where $\widetilde{t}$

is the posterior replicates of the test statistic t. The final ppp-value is the posterior mean of $p(\mathbf{W}^{\mathrm{mis}})$ given the data, which is $\mathrm{ppp}(H_0^c) = \int p(\mathbf{W}^{\mathrm{mis}}) \Pi(d\mathbf{W}^{\mathrm{mis}}|\mathbf{O}^{\mathrm{obs}})$.

Operationally, the key is to obtain the posterior draws of $\mathbf{W}^{\mathrm{mis}}$. Under $\mathbf{Y}(1) = \mathbf{Y}(0) = \mathbf{Y}^{\mathrm{obs}}$, the observed outcomes $\mathbf{Y}^{\mathrm{obs}}$ can be viewed as *pretreatment covariates* under monotonicity, exclusion restriction and H_0^c and, therefore, we do not need to model the outcome. Assume that first part of the joint model $\mathrm{Pr}(\mathbf{Z}|\mathbf{X}, \mathbf{U}, \mathbf{Y}^{\mathrm{obs}}) = \mathrm{Pr}(\mathbf{Z}|\mathbf{X})$ is known, as often guaranteed by the design of experiments. Therefore, we need only to model the second part $\mathrm{Pr}(\mathbf{U}|\mathbf{X}, \mathbf{Y}^{\mathrm{obs}}, \theta)$ and impose a prior $\mathrm{Pr}(\theta)$. Assuming exchangeability of the units, we define $\pi_u(X_i, Y_i^{\mathrm{obs}}, \theta) = \mathrm{Pr}(U_i = u|X_i, Y_i^{\mathrm{obs}}, \theta)$ for $u = \mathrm{c}, \mathrm{n}$ and a, which, for example, can be modeled as multinomial logistic. We can obtain the posterior distribution $\Pi(\mathbf{W}^{\mathrm{mis}}|\mathbf{O}^{\mathrm{obs}})$ as a byproduct of the iteration of the data augmentation algorithm for sampling $(\mathbf{W}^{\mathrm{mis}}, \theta)$, by discarding the posterior draws of θ. In particular, given $\mathbf{W}^{\mathrm{mis}}$, we know $\mathbf{U}$ and, therefore, we can use $(\mathbf{U}, \mathbf{X}, \mathbf{Y}^{\mathrm{obs}})$ to obtain a posterior draw of θ, for example, according to the posterior distribution of a multinomial logistic model. Given θ, we can impute $\mathbf{W}^{\mathrm{mis}}$ as follows: for a unit with $(Z_i = 1, W_i^{\mathrm{obs}} = 1)$, draw $W_i(0)$ from Bernoulli with probability $\pi_{\mathrm{a}}(X_i, Y_i^{\mathrm{obs}}, \theta)/\{\pi_{\mathrm{a}}(X_i, Y_i^{\mathrm{obs}}, \theta) + \pi_{\mathrm{c}}(X_i, Y_i^{\mathrm{obs}}, \theta)\}$; for a unit with $(Z_i = 1, W_i^{\mathrm{obs}} = 0)$, set $W_i(0) = 0$; for a unit with $(Z_i = 0, W_i^{\mathrm{obs}} = 1)$, set $W_i(1) = 1$; for a unit with $(Z_i = 0, W_i^{\mathrm{obs}} = 0)$, draw $W_i(1)$ from Bernoulli with probability $\pi_{\mathrm{c}}(X_i, Y_i^{\mathrm{obs}}, \theta)/\{\pi_{\mathrm{c}}(X_i, Y_i^{\mathrm{obs}}, \theta) + \pi_{\mathrm{n}}(X_i, Y_i^{\mathrm{obs}}, \theta)\}$. Rubin (1998) described this strategy for the one-sided noncompliance with $W_i(0) = 0$ for all units under a completely randomized experiment; his simulation studies for the case with a binary outcome without $\mathbf{X}$ showed meaningful power gains in some scenarios.

4.4 Extension

In Section 4.3, we focused on obtaining the p-value based on a test statistic $t(\mathbf{O}^{\mathrm{obs}})$, a function of the observed data only. In general, we can use a *discrepancy variable* $t(\mathbf{O}^{\mathrm{obs}}, \mathbf{O}^{\mathrm{mis}})$, which can be a function of both the observed and missing data, to obtain the p-value (Meng, 1994, Gelman, Meng and Stern, 1996). In both cases, the p-value is a function of $\mathbf{O}^{\mathrm{mis}}$.

The above discussion of the Fisher randomization inference applies naturally to randomized studies in which $\mathrm{Pr}(\mathbf{W}|\mathbf{X})$ is known and determined by the designers of the experiments. Complications arise in observational studies. If Assumption 1 holds and X is

discrete, then we can use the estimated propensity score $\widehat{\mathrm{Pr}}(\mathbf{W}|\mathbf{X})$ to simulate the treatment $\mathbf{W}$, or equivalently conduct conditional randomization test as if the data come from a stratified completely randomized experiment. With continuous or multidimensional X, we need to model $\mathrm{Pr}(\mathbf{W}|\mathbf{X})$. Rosenbaum (1984a) proposed a conditional randomization test given the sufficient statistics of a logistic model for $\mathrm{Pr}(\mathbf{W}|\mathbf{X})$. Rubin (2007, 2008) and Imbens and Rubin (2015) suggested stratifying on the estimated propensity scores to approximate completely randomized experiments within strata, possibly followed by a Fisherian analysis.

Although we focused on obtaining p-values from Fisher randomization tests, we can invert a sequence of tests to obtain confidence sets of parameters of interest. With a few exceptions, this is often conducted under an additional assumption of constant treatment effect (Rosenbaum, 2002b). However, with nonsharp null hypotheses, simultaneously imputing missing potential outcomes and inverting a sequence of tests seems inferior to both the Frequentist and Bayesian perspectives discussed before, which focus on point and interval estimation directly.

Moreover, it is important to evaluate the frequency properties of the ppp-values. Under the classical Frequentist evaluation in which the parameters are fixed constants, the Fisherian p-values are *exact* when the null hypotheses are sharp, but they are often conservative when the null hypotheses are not sharp (Robins, van der Vaart and Ventura, 2000). However, alternative frequency evaluations often give different conclusions. Meng (1994) considered the *prior predictive evaluation*, in which the model parameters are generated from proper prior distributions and the data are then generated conditional on the model parameters. Under Meng's (1994) scheme, some ppp-values may be anti-conservative. Rubin (1998) considered the *posterior predictive evaluation*, in which the model parameters and missing data are generated conditional on the observed data and the null hypotheses. Under Rubin's (1998) scheme, the ppp-values have exact frequency properties.

As a final remark on the Fisherian perspective, neither did Fisher formally use potential outcomes nor did he agree with using Bayes' theorem for statistical inference (Fisher, 1935). Rubin (1980) reformulated the randomization test using Neyman's (1990) potential outcomes notation, and extended it by combining p-values with Bayesian techniques. Alternatively, assuming that the units are independent and identically draws from a super population, Hoeffding (1952) and Chung

and Romano (2013) stated the "sharp" null hypothesis as $Y_i(1) \sim Y_i(0)$, that is, the treated and control potential outcomes have the same distribution, and interpreted randomization tests as *permutation tests*. This interpretation works well under the sharp null hypotheses for completely randomized experiments, but is less straightforward for nonsharp null hypotheses with nuisance parameters. The super-population version of the analysis in Section 4.3 remains an open question.

5. DISCUSSION

We have reviewed a wide range of causal inference methods that have analogues in missing data analysis under three modes of inference. Although a comprehensive review of all relevant topics is beyond the scope of this paper given, the vast literature in both causal inference and missing data, we regard the following important areas merit further attention.

5.1 Partially Identified Parameters and Bounds

In the Bayesian perspective, we commented on the identifiability issue due to the fundamental problem of causal inference and recommended to use a transparent parametrization strategy (see Section 3.1), which also applies to the Fisherian perspective. We now comment more on the identifiability issue in the Frequentist perspective, which often involves bounding the parameter of interest by the distributions of the observables. Cochran (1953) derived bounds for nonignorable missing data in surveys, and Manski (1990) obtained more fruitful results for both missing data and causal inference problems.

The first important class of partially identified parameters depend on the association between the potential outcomes. For example, the parameters

$$\delta_1 = \Pr(Y(1) \geq Y(0)), \quad \delta_2 = \Pr(Y(1) > Y(0))$$

measure the probability that the treatment is not worse than the control and the probability that the treatment is better than the control, respectively. Lu, Ding and Dasgupta (2015) emphasize that for ordinal outcomes, δ_1 and δ_2 are well defined even though $\tau_i = Y_i(1) - Y_i(0)$ is not. In general, we are interested in the distribution of the treatment effect $\Delta(c) = \Pr(Y_i(1) - Y_i(0) \leq c)$. Without imposing further assumptions, we can only bound parameters such as δ_1, δ_2 and $\Delta(c)$ by the marginal distributions $\Pr(Y_i(1) \leq y_1)$ and $\Pr(Y_i(0) \leq y_0)$.

Inferring such parameters is arguably more challenging than most standard missing data problems. It is

also an example where some theoretical development in causal inference lends to research on missing data. Specifically, a small but growing literature is on the so-called "misaligned missing data" problem (e.g., Ding and Song, 2016), where some variables cannot be observed simultaneously, similar to the joint potential outcomes $\mathbf{Y}(1)$ and $\mathbf{Y}(0)$. Ding and Song (2016) made the connection and borrowed some results from causal inference to address the problem (Fan, Guerre and Zhu, 2017).

Inference involving simultaneous counterfactual potential outcomes is controversial for some researchers who were only willing to model observables. For example, Dawid (2000) advocated "causal inference without counterfactuals" through a decision theoretic perspective. Interestingly, despite his original critique, Dawid, Musio and Murtas (2017) invoked counterfactuals to define the "probability of causation (PC)." For example, if a unit responds to the treatment, then

$$\text{PC} = \Pr\big(Y_i(0) = 0 | W_i = 1, Y_i(1) = 1\big)$$

is the probability that the treatment is effective. The parameter PC obviously involves the joint values of the counterfactuals. Indeed, such joint counterfactuals are often unavoidable in defining certain causal estimands, such as δ_1, δ_2 and $\Delta(c)$. Heckman, Lopes and Piatek (2014) provided more discussions on this issue.

Another important class of partially identified parameters arises from principal stratification. If we do not invoke monotonicity, exclusion restriction or parametric models, in general we cannot identify the means of the potential outcomes within the latent principal strata. Zhang and Rubin (2003) derived large-sample bounds of the causal parameters based on the observed data, followed by Cheng and Small (2006), Grilli and Mealli (2008), Imai (2008), among others.

Unfortunately, the bounds are often too wide to be useful in practice, and additional information is often required to sharpen them (Mattei and Mealli, 2011, Yang and Small, 2016). It is also nontrivial to construct confidence intervals for the bounds or for the parameters themselves, because the bounds often correspond to nonsmooth operations of the observed data distribution and, therefore, the standard delta method or the bootstrap may not apply (Andrews, 2000, Fan and Park, 2010).

5.2 Nonignorable Assignment Mechanisms and Sensitivity Analysis

Our discussion is limited to ignorable (or unconfounded) assignment mechanisms. However, in observational studies the assignment mechanism is generally unknown, and it is the norm rather exception that

the untestable unconfoundedness assumption (2) is violated to a certain degree. The critical reliance on unconfoundedness in causal inference is similar to the dependence on the vulnerable MAR assumption in missing data analysis. In causal inference, the standard approach is to conduct sensitivity analysis around unconfoundedness. The main idea is to examine the causal estimates from the same inferential procedure given an unmeasured covariate whose confounding effects (i.e., degree of violation to unconfoundedness) are encoded as sensitivity parameters and are varied within a realistic range by the analyst. Cornfield et al. (1959) first started this school of thoughts, whose method was directly extended in Ding and Vander-Weele (2016). Rosenbaum and Rubin (1983a) broadened this approach by parametrically modeling the distributions of the outcome and the treatment conditional on the unmeasured confounder. Semi and nonparametric versions of this method have been developed (e.g., Rosenbaum, 1987, Imbens, 2003, Ichino, Mealli and Nannicini, 2008, Robins, Rotnitzky and Scharfstein, 2000), and extensions to principal stratification are also available (e.g., Schwartz, Li and Reiter, 2012, Mercatanti and Li, 2017, Ding and Lu, 2017).

Sensitivity analysis is also extensively conducted in the context of nonignorable missing data, but with different emphasis and implementation than that in causal inference. Specifically, the two standard models for nonignorable missing data are the *selection* models and the *pattern-mixture* models (Little and Rubin, 2002, Chapter 15); in both models, the missing data mechanism is directly modeled and estimated from the data, together with the outcome model. Because sensitivity to model specification is a serious scientific problem for both selection and pattern-mixture models, in real applications it is prudent to consider a variety of missing data models, rather than to rely exclusively on one model. One attractive direction for both causal inference and missing data with nonignorable assignment/missing data mechanisms is to utilize external data sources to augment the analysis, which requires considerable efforts in the study design.

5.3 Unintentional Missing Data

We have mainly focused on the intentional missing data (i.e., missing potential outcomes) in causal inference. Unintentional missing data are prevalent in observational studies and there is a growing literature on drawing causal inference from incomplete data. A straightforward approach consists of two independent steps: first impute the missing data and then draw causal inferences from the imputed complete data (e.g., Mitra and Reiter, 2011). However, how the missing values are imputed may have a nontrivial impact on the subsequent causal analysis (e.g., Mitra and Reiter, 2016). In particular, missing data in covariates (Rosenbaum and Rubin, 1984, Ding and Geng, 2014, Yang, Wang and Ding, 2017), treatment (Molinari, 2010, Mebane and Poast, 2013, Zhang et al., 2016), and outcomes (Frangakis and Rubin, 1999, Chen, Geng and Zhou, 2009, Mattei, Mealli and Pacini, 2014, Liublinska and Rubin, 2014) are often of different nature and may require different handling. The problem is even more challenging in complex settings such as post-treatment variables (e.g., Mercatanti, 2004, Mealli et al., 2004). More research in understanding such an impact would be valuable to practice.

5.4 Software

Open-source and user-friendly software packages are crucial for bridging theory and practice, and deserve much effort and investment from methodologists. In statistics, authors are increasingly releasing R packages implementing their methods. This is particularly important for causal inference methods, which are widely used in many substantive disciplines including medicine, policy, psychology, social sciences and others. We list a few most popular packages here: twang (Ridgeway et al., 2017) provides functions for propensity score estimating (via generalized boosted models) and weighting methods; packages Matching (Sekhon, 2011) and MatchIt (Ho et al., 2011) provide functions for matching methods. These packages can be used in combination with other R packages, for example, for hierarchial models (lme4), Bayesian modeling (rjags), to apply to more complex problems. Besides R, Stan (Carpenter et al., 2017) is new open-source probabilistic programming language particulary suitable for advanced statistical modeling and computation; Stan has attracted much interest from both academia and industry recently and can be used for implementing many causal inference methods discussed in this review.

REFERENCES

ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267. MR2194325

ABADIE, A. and IMBENS, G. (2011). Bias corrected matching estimators for average treatment effects. *J. Bus. Econom. Statist.* **29** 1–11. MR2789386

ANDREWS, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68** 399–405. MR1748009

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455. MR3042387

ANGRIST, J. D. and PISCHKE, J.-S. (2008). *Mostly Harmless Econometrics*: *An Empiricist's Companion*. Princeton Univ. Press, Princeton, NJ.

ATHEY, S. and IMBENS, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. Available at arXiv:1504.01132.

ATHEY, S., IMBENS, G. W. and WAGER, S. (2018). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* To appear. Available at https://arxiv.org/abs/1604.07125.

ATHEY, S., IMBENS, G., PHAM, T. and WAGER, S. (2017). Estimating average treatment effects: Supplementary analyses and remaining challenges. *Am. Econ. Rev.* **107** 278–281.

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983

BELLONI, A., CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and HANSEN, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85** 233–298. MR3611771

BICKEL, P. J. and DOKSUM, K. A. (2015). *Mathematical Statistics*: *Basic Ideas and Selected Topics*, *Volume I*, 2nd ed. CRC Press, Boca Raton, FL. MR3445928

BLONIARZ, A., LIU, H., ZHANG, C.-H., SEKHON, J. S. and YU, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proc. Natl. Acad. Sci. USA* **113** 7383–7390. MR3531136

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. and RIDDELL, A. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76** 1–32.

CHAPIN, F. S. (1947). *Experimental Designs in Sociological Research*. Harper, New York.

CHEN, H., GENG, Z. and ZHOU, X.-H. (2009). Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data. *Biometrics* **65** 675–682. MR2649840

CHENG, J. and SMALL, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 815–836. MR2301296

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2016). Double/debiased machine learning for treatment and causal parameters. Preprint. Available at arXiv:1608.00060.

CHIB, S. and JACOBI, L. (2016). Bayesian fuzzy regression discontinuity analysis and returns to compulsory schooling. *J. Appl. Econometrics* **31** 1026–1047. MR3556650

CHUNG, E. and ROMANO, J. P. (2013). Exact and asymptotically robust permutation tests. *Ann. Statist.* **41** 484–507. MR3099111

COCHRAN, W. G. (1953). *Sampling Techniques*, 1st ed. Wiley, New York. MR0054199

COCHRAN, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics* **13** 261–281. MR0090952

COCHRAN, W. G. (2007). *Sampling Techniques*, 3rd ed. Wiley, New York.

CORNFIELD, J., HAENSZEL, W., HAMMOND, E. et al. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* **22** 173–203.

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. MR2482144

DAWID, A. P. (2000). Causal inference without counterfactuals. *J. Amer. Statist. Assoc.* **95** 407–424. MR1803167

DAWID, A. P. MUSIO, M. and MURTAS, R. (2017). The probability of causation. *Law*, *Probability and Risk* **16** 163–179.

DEMPSTER, A., LAIRD, N. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39** 1–38. MR0501537

DING, P. (2014). Three occurrences of the hyperbolic-secant distribution. *Amer. Statist.* **68** 32–35. MR3303831

DING, P. and DASGUPTA, T. (2016). A potential tale of two-by-two tables from completely randomized experiments. *J. Amer. Statist. Assoc.* **111** 157–168. MR3494650

DING, P., FELLER, A. and MIRATRIX, L. (2016). Randomization inference for treatment effect variation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 655–671. MR3506797

DING, P. and GENG, Z. (2014). Identifiability of subgroup causal effects in randomized experiments with nonignorable missing covariates. *Stat. Med.* **33** 1121–1133. MR3247784

DING, P. and LU, J. (2017). Principal stratification analysis using principal scores. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 757–777. MR3641406

DING, P. and VANDERWEELE, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27** 368–377.

DING, P., GENG, Z., YAN, W. and ZHOU, X.-H. (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *J. Amer. Statist. Assoc.* **106** 1578–1591. MR2896858

DING, W. and SONG, P. X.-K. (2016). EM algorithm in Gaussian copula with missing data. *Comput. Statist. Data Anal.* **101** 1–11. MR3504831

ELLIOTT, M., RAGHUNATHAN, T. and LI, Y. (2010). Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics* **11** 353–372.

FAN, Y., GUERRE, E. and ZHU, D. (2017). Partial identification of functionals of the joint distribution of "potential outcomes". *J. Econometrics* **197** 42–59. MR3598644

FAN, Y. and PARK, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory* **26** 931–951.

FELLER, A., GREIF, E., MIRATRIX, L. and PILLAI, N. (2016). Principal stratification in the twilight zone: Weakly separated components in finite mixture models. Preprint. Available at arXiv:1602.06595.

FIRTH, D. and BENNETT, K. E. (1998). Robust models in probability sampling. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 3–21. MR1625672

FISHER, R. A. (1935). *The Design of Experiments*, 1st ed. Oliver and Boyd, Edinburgh.

FRANGAKIS, C. and RUBIN, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86** 365–378. MR1705410

FRANGAKIS, C. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039

FRUMENTO, P., MEALLI, F., PACINI, B. and RUBIN, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *J. Amer. Statist. Assoc.* **107** 450–466. MR2980057

FRUMENTO, P., MEALLI, F., PACINI, B. and RUBIN, D. B. (2016). The fragility of standard inferential approaches in principal stratification models relative to direct likelihood approaches. *Stat. Anal. Data Min.* **9** 58–70. MR3465093

GALLOP, R., SMALL, D., LIN, J., ELLIOT, M., JOFFE, M. and HAVE, T. T. (2009). Mediation analysis with principal stratification. *Stat. Med.* **28** 1108–1130. MR2662200

GELFAND, A. and SMITH, A. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sinica* **6** 733–807. MR1422404

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2027492

GILBERT, P. and HUDGENS, M. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64** 1146–1154. MR2522262

GRAHAM, B. S., DE XAVIER PINTO, C. C. and EGEL, D. (2012). Inverse probability tilting for moment condition models with missing data. *Rev. Econ. Stud.* **79** 1053–1079. MR2986390

GRILLI, L. and MEALLI, F. (2008). Nonparametric bounds on the causal effect of university studies on job opportunities using principal stratification. *J. Educ. Behav. Stat.* **33** 111–130.

GUSTAFSON, P. (2009). What are the limits of posterior distributions arising from nonidentified models, and why should we care? *J. Amer. Statist. Assoc.* **104** 1682–1695. MR2750585

GUSTAFSON, P. (2015). *Bayesian Inference for Partially Identified Models*: *Exploring the Limits of Limited Data*. CRC Press, Boca Raton, FL.

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331. MR1612242

HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20** 25–46.

HÁJEK, J. (1971). Comment on a paper by D. Basu. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 236. Holt, Rinehart and Winston, Toronto.

HECKMAN, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161. MR0518832

HECKMAN, J., LOPES, H. and PIATEK, R. (2014). Treatment effects: A Bayesian perspective. *Econometric Rev.* **33** 36–67. MR3170840

HIRANO, K. and IMBENS, G. W. (2004). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. 73–84. Wiley, Chichester. MR2134803

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *J. Stat. Softw.* **42** 1–28.

HOEFFDING, W. (1952). The large-sample power of tests based on permutations of observations. *Ann. Math. Stat.* **23** 169–192. MR0057521

HOLLAND, P. (1986). Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

HORVITZ, D. and THOMPSON, D. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* **1** 221–233. Univ. California Press, Berkeley, CA. MR0216620

ICHINO, A., MEALLI, F. and NANNICINI, T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *J. Appl. Econometrics* **23** 305–327. MR2420362

IMAI, K. (2008). Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". *Statist. Probab. Lett.* **78** 144–149. MR2382067

IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 243–263. MR3153941

IMAI, K. and VAN DYK, D. (2004). Causal treatment with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* **99** 854–866. MR2090918

IMBENS, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710. MR1789821

IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* **93** 126–132.

IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.

IMBENS, G. W. and ANGRIST, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–476.

IMBENS, G. W. and RUBIN, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.* **25** 305–327. MR1429927

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference for Statistics*, *Social*, *and Biomedical Sciences*: *An Introduction*. Cambridge Univ. Press, New York. MR3309951

KANG, J. D. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458

LI, X. and DING, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.* **112** 1759–1769. MR3750897

LI, F., MATTEI, A. and MEALLI, F. (2015). Evaluating the causal effect of university grants on student dropout: Evidence from a regression discontinuity design using principal stratification. *Ann. Appl. Stat.* **9** 1906–1931. MR3456358

LI, F., MORGAN, K. and ZASLAVSKY, A. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* To appear. Available at https://doi.org/10.1080/01621459.2016.1260466.

LI, F., BACCINI, M., MEALLI, F., ZELL, E. R., FRANGAKIS, C. E. and RUBIN, D. B. (2014). Multiple imputation by ordered monotone blocks with application to the anthrax vaccine research program. *J. Comput. Graph. Statist.* **23** 877–892. MR3224660

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *Ann. Appl. Stat.* **7** 295–318. MR3086420

LINDLEY, D. V. (1972). *Bayesian Statistics*: *A Review*. SIAM, Philadelphia, PA. MR0329081

LITTLE, R. J. (1988). Missing-data adjustments in large surveys. *J. Bus. Econom. Statist.* **6** 287–296.

LITTLE, R. and AN, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statist. Sinica* **14** 949–968. MR2089342

LITTLE, R. J. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley-Interscience, Hoboken, NJ. MR1925014

LIUBLINSKA, V. and RUBIN, D. B. (2014). Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Stat. Med.* **33** 4170–4185. MR3267402

LU, J., DING, P. and DASGUPTA, T. (2015). Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. Preprint. Available at arXiv:1507.01542.

LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **23** 2937–2960.

MANSKI, C. F. (1990). Nonparametric bounds on treatment effects. *Am. Econ. Rev.* **80** 319–323.

MATTEI, A. and MEALLI, F. (2011). Augmented designs to assess principal strata direct effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 729–752. MR2867456

MATTEI, A., MEALLI, F. and PACINI, B. (2014). Identification of causal effects in the presence of nonignorable missing outcome values. *Biometrics* **70** 278–288. MR3258033

MEALLI, F. and RUBIN, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **102** 995–1000. MR3431570

MEALLI, F., IMBENS, G. W., FERRO, S. and BIGGERI, A. (2004). Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics* **5** 207–222.

MEBANE, W. R. JR and POAST, P. (2013). Causal inference without ignorability: Identification with nonrandom assignment and missing treatment data. *Polit. Anal.* **21** 233–251.

MENG, X.-L. (1994). Posterior predictive *p*-values. *Ann. Statist.* **22** 1142–1160. MR1311969

MERCATANTI, A. (2004). Analyzing a randomized experiment with imperfect compliance and ignorable conditions for missing data: Theoretical and computational issues. *Comput. Statist. Data Anal.* **46** 493–509. MR2061964

MERCATANTI, A. and LI, F. (2014). Do debit cards increase household spending? Evidence from a semiparametric causal analysis of a survey. *Ann. Appl. Stat.* **8** 2405–2508. MR3292506

MERCATANTI, A. and LI, F. (2017). Do debit cards decrease cash demand? Causal inference and sensitivity analysis using principal stratification. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 759–776. MR3670416

MIRATRIX, L. W., SEKHON, J. S. and YU, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 369–396. MR3021392

MITRA, R. and REITER, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Stat. Med.* **30** 627–641. MR2767460

MITRA, R. and REITER, J. P. (2016). A comparison of two methods of estimating propensity scores after multiple imputation. *Stat. Methods Med. Res.* **25** 188–204.

MOLINARI, F. (2010). Missing treatments. *J. Bus. Econom. Statist.* **28** 82–95. MR2650602

MURRAY, J. S. and REITER, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *J. Amer. Statist. Assoc.* **111** 1466–1479. MR3601702

NEWEY, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79** 147–168. MR1457700

NEYMAN, J. (1935). Statistical problems in agricultural experimentation. *Suppl. J. R. Stat. Soc.* **2** 107–180.

NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. MR1092986

NOLEN, T. L. and HUDGENS, M. G. (2011). Randomization-based inference within principal strata. *J. Amer. Statist. Assoc.* **106** 581–593. MR2847972

QIN, J. (2017). *Biased Sampling*, *Over-Identified Parameter Problems and Beyond*. Springer, Singapore. MR3675467

RICHARDSON, T. S., EVANS, R. J. and ROBINS, J. M. (2010). Transparent parameterizations of models for potential outcomes. In *Bayesian Statistics* 9 569–610. Oxford Univ. Press, Oxford.

RIDGEWAY, G., MCCAFFREY, D., MORRAL, A., GRIFFIN, B. A. and BURGETTE, L. (2017). twang: Toolkit for Weighting and Analysis of Nonequivalent Groups. R package version 1.5.

ROBINS, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Math. Modelling* **7** 1393–1512. MR0877758

ROBINS, J. M. and RITOV, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Stat. Med.* **16** 285–319.

ROBINS, J. M., ROTNITZKY, A. and SCHARFSTEIN, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology*, *the Environment*, *and Clinical Trials* 1–94. Springer, New York. MR1731681

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90** 106–121. MR1325118

ROBINS, J. M., VAN DER VAART, A. and VENTURA, V. (2000). Asymptotic distribution of *p* values in composite null models. *J. Amer. Statist. Assoc.* **95** 1143–1156. MR1804240

ROSENBAUM, P. R. (1984a). Conditional permutation tests and the propensity score in observational studies. *J. Amer. Statist. Assoc.* **79** 565–574. MR0763575

ROSENBAUM, P. R. (1984b). The consquences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **147** 656–666.

ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. MR0885915

ROSENBAUM, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. MR1962487

ROSENBAUM, P. R. (2002b). *Observational Studies*, 2nd ed. Springer, New York. MR1899138

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York. MR2561612

ROSENBAUM, P. R. and RUBIN, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **45** 212–218.

ROSENBAUM, P. R. and RUBIN, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.

RUBIN, D. B. (1975). Bayesian inference for causality: The role of randomization. In *Proceedings of the Social Statistics Section of the American Statistical Association* 233–239.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

RUBIN, D. B. (1977). Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics* **2** 1–26.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–324.

RUBIN, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *Ann. Statist.* **12** 1151–1172.

RUBIN, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econom. Statist.* **4** 87–94.

RUBIN, D. B. (1998). More powerful randomization-based *p*-values in double-blind trials with non-compliance. *Stat. Med.* **17** 371–385.

RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071

RUBIN, D. B. (2006a). Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Statist. Sci.* **91** 299–321. MR2339125

RUBIN, D. B. (2006b). *Matched Sampling for Causal Effects*. Cambridge Univ. Press, Cambridge. MR2307965

RUBIN, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat. Med.* **26** 20–36. MR2312697

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–840. MR2516795

SCHARFSTEIN, D., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478

SCHWARTZ, S., LI, F. and REITER, J. P. (2012). Sensitivity analysis for unmeasured confounding in principal stratification settings with binary variables. *Stat. Med.* **31** 949–962. MR2913871

SEAMAN, S., GALATI, J., JACKSON, D. and CARLIN, J. (2013). What is meant by "missing at random"? *Statist. Sci.* **28** 257–268. MR3112409

SEKHON, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *J. Stat. Softw.* **42** 1–52.

STUART, E. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812

TANNER, M. and WONG, W. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–540. MR0898357

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LU, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Stat. Med.* **27** 4658–4677. MR2528575

TUKEY, J. W. (1993). Tightening the clinical trial. *Controlled Clinical Trials* **14** 266–285.

VAN BUUREN, S. (2012). *Flexible Imputation of Missing Data*. CRC press, Boca Raton, FL.

VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York. MR2867111

VANDERWEELE, T. (2008). Simple relations between principal stratification and direct and indirect effects. *Statist. Probab. Lett.* **78** 2957–2962. MR2516810

WAGER, S., DU, W., TAYLOR, J. and TIBSHIRANI, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proc. Natl. Acad. Sci. USA* **113** 12673–12678. MR3576188

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838. MR0575027

YANG, S. and DING, P. (2018). Asymptotic causal inference with observational studies trimmed by the estimated propensity scores. *Biometrika*. To appear. Available at https://arxiv.org/abs/1604.07125.

YANG, F. and SMALL, D. S. (2016). Using post-outcome measurement information in censoring-by-death problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 299–318. MR3453657

YANG, S., WANG, L. and DING, P. (2017). Nonparametric identification of causal effects with confounders subject to instrumental missingness. Preprint. Available at arXiv:1702.03951.

ZHANG, G. and LITTLE, R. J. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics* **65** 911–918. MR2649864

ZHANG, J. and RUBIN, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *J. Educ. Behav. Stat.* **28** 353–358.

ZHANG, J., RUBIN, D. B. and MEALLI, F. (2009). Likelihood-based analysis of the causal effects of job-training programs using principal stratification. *J. Amer. Statist. Assoc.* **104** 166–176.

ZHANG, Z., LIU, W., ZHANG, B., TANG, L. and ZHANG, J. (2016). Causal inference with missing exposure information: Methods and applications to an obstetric study. *Stat. Methods Med. Res.* **25** 2053–2066. MR3553326

ZHOU, J., ZHANG, Z., LI, Z. and ZHANG, J. (2015). Coarsened propensity scores and hybrid estimators for missing data and causal inference. *Int. Stat. Rev.* **83** 449–471. MR3429286

ZIGLER, C. and BELIN, T. (2012). A Bayesian approach to improved estimation of causal effect predictiveness for a principal surrogate endpoint. *Biometrics* **68** 922–932. MR3055197

ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* **110** 910–922. MR3420672