PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Review

Cite this article: Li F, Ding P, Mealli F. 2023 Bayesian causal inference: a critical review. *Phil. Trans. R. Soc. A* **381**: 20220153. https://doi.org/10.1098/rsta.2022.0153

Received: 28 June 2022
Accepted: 23 October 2022

One contribution of 16 to a theme issue 'Bayesian inference: challenges, perspectives, and prospects'.

Subject Areas:
statistics

Keywords:
causal inference, design, ignorability, potential outcomes, propensity score

Author for correspondence:
Fan Li
e-mail: fl35@duke.edu

Bayesian causal inference: a critical review

Fan Li[1], Peng Ding[2] and Fabrizia Mealli[3]

[1]Duke University, Durham, NC, USA
[2]University of California, Berkeley, CA, USA
[3]University of Florence and EUI, Florence, Italy

FL, 0000-0002-0390-3673

This paper provides a critical review of the Bayesian perspective of causal inference based on the potential outcomes framework. We review the causal estimands, assignment mechanism, the general structure of Bayesian inference of causal effects and sensitivity analysis. We highlight issues that are unique to Bayesian causal inference, including the role of the propensity score, the definition of identifiability, the choice of priors in both low- and high-dimensional regimes. We point out the central role of covariate overlap and more generally the design stage in Bayesian causal inference. We extend the discussion to two complex assignment mechanisms: instrumental variable and time-varying treatments. We identify the strengths and weaknesses of the Bayesian approach to causal inference. Throughout, we illustrate the key concepts via examples.

This article is part of the theme issue 'Bayesian inference: challenges, perspectives, and prospects'.

1. Introduction

Causality has long been central to the human philosophical debate and scientific pursuit. There are many relevant questions, e.g. the philosophical meaning of causation or deducing the causes of a given phenomenon. Among these questions, statistics—which concerns measurements—arguably can contribute the most to the question of measuring the effects of causes. Statistics infers associations between variables. Even though the research questions in many statistics-based studies are causal in nature, a first lesson in elementary statistics is that *association does not imply causation*. Distinguishing between causation and spurious association between various events is a challenging task in science. Broadly speaking, statistical causal inference is about building a framework that (i)

THE ROYAL SOCIETY PUBLISHING

defines causal effects under general scenarios, (ii) specifies assumptions under which one can identify causation from association and (iii) assesses the sensitivity to the causal assumptions and finds ways to mitigate.

A mainstream statistical framework for causal inference is the potential outcomes framework [1–3]. Under this framework, following the dictum '*no causation without manipulation*' [3], a *cause* is a pre-specified *treatment* or *intervention* that is at least hypothetically manipulable. A typical causal question is 'would an individual have a better outcome had he taken treatment A versus treatment B?' Causal effects are defined as comparisons of potential outcomes, also known as counterfactuals, under different treatment conditions for the same units. The main hurdle to interpreting the association between the treatment and the outcome as a causal effect is confounding, i.e. the presence of factors that are associated with both the treatment and the outcome. For example, patients with worse health conditions may be more likely to obtain a beneficial medical treatment; then directly comparing the outcomes of the treated and control patients, without adjusting for the difference in their baseline health conditions, would bias the causal comparisons and mistakenly conclude that the treatment is harmful. Randomized experiments, known as *A/B testing* in industry or randomized controlled trials in medicine, are the gold standard for casual inference by eliminating confounding via randomization. But modern causal inference has increasingly relied on observational data. The potential outcomes framework provides the basis for identifying and estimating causal effects—quantities defined based on counterfactuals—from the factual data in the presence of confounding, using randomized or observational data. This framework is applicable to a wide range of problems in many disciplines and has been increasingly adopted in the area of machine learning. Other frameworks for causal inference, including the causal diagram [4] and invariant prediction [5], are beyond the scope of this review.

There are three primary inferential approaches within the potential outcomes framework [6]: Fisherian randomization test, Neymanian repeated-sampling evaluation and Bayesian inference. The first two approaches belong to the Frequentist paradigm and have been dominant, with many popular tools such as propensity scores, matching and weighting. The Bayesian approach has several established advantages for general statistical analysis, including automatic uncertainty quantification, coherently incorporating prior knowledge, and offering a rich collection of advanced models for complex data. As causal studies increasingly involve real-world big data, there has been a recent surge of research in Bayesian inference of causal effects [7–11], but there is no comprehensive appraisal of the current state of the research. This paper aims to fill this gap. Due to the space limit, we do not intend to provide a catalogue of the existing research on this topic, but rather discuss the big picture of why and how to conduct Bayesian causal inference in general settings. We emphasize the unique questions, challenges and opportunities that the Bayesian approach brings to causal inference. We hope this review can stimulate broader and deeper cross-fertilization between causal inference and Bayesian analysis.

Section 2 introduces the preliminaries of the potential outcomes framework, and briefly discusses several Frequentist methods to causal inference. Section 3 outlines the general structure of Bayesian causal inference, focusing on ignorable treatment assignments at one time point. Section 4 discusses model specification and implications in high-dimensional settings. Section 5 reviews the role and various uses of the propensity score in Bayesian causal inference. Section 6 outlines sensitivity analysis in observational studies. Section 7 describes two complex assignment mechanisms: instrumental variable and time-varying treatments. Section 8 concludes.

2. Estimands, identification and frequentist estimation

To convey the main ideas, we focus on the case with a binary treatment at one time period, which can be readily extended to multiple treatments and multiple time points. Consider a sample of units drawn from a target population, indexed by $i \in \{1, \ldots, N\}$. Each unit can potentially be assigned to one of two treatment levels z, with $z = 1$ for the active treatment and $z = 0$ for the control. Let $Z_i (= z)$ be the binary variable indicating unit i's observed treatment status. For unit i,

a vector of p covariates X_i are observed before the treatment, and an outcome Y_i is observed after the treatment. A confounder is a pre-treatment variable that is associated with both the treatment and the outcome; it can be observed, as a subset of the covariates X_i, or unobserved. Below we use covariates and confounders interchangeably. We use the $A \perp\!\!\!\perp B \mid C$ notation to denote conditional independence between two variables A and B given variable C [12]. We also use the bold font to indicate a vector consisting of the corresponding variables for the N units, e.g. $\mathbf{Z} = (Z_1, \ldots, Z_N)'$.

We maintain the standard stable unit treatment value assumption (SUTVA) [13], namely, there is (i) no different version of a treatment, and (ii) no interference in the sense that one unit's potential outcomes are not affected by other units' treatment assignment. Under SUTVA, each unit i has two potential outcomes: $Y_i(1)$ and $Y_i(0)$. Causal effects are contrasts of potential outcomes under different treatment conditions for the same set of units. The individual treatment effect (ITE) for unit i is $\tau_i = Y_i(1) - Y_i(0)$. Averaging τ_i over a sample we obtain the sample average treatment effect (SATE): $\tau^{\mathrm{S}} \equiv N^{-1} \sum_{i=1}^{N} \tau_i$. Furthermore, the conditional average treatment effect (CATE) is the average of the individual treatment effect of all units with the covariate value x:

$$\tau(x) \equiv \mathbb{E}\{Y_i(1) - Y_i(0) \mid X_i = x\} = \mu_1(x) - \mu_0(x), \tag{2.1}$$

where $\mu_z(x) \equiv \mathbb{E}\{Y_i(z) \mid X_i = x\}$ for $z = 0, 1$. Averaging τ_i or $\tau(X_i)$ over a target population gives the population average treatment effect (PATE):

$$\tau^{\mathrm{P}} \equiv \mathbb{E}\{Y_i(1) - Y_i(0)\} = \mathbb{E}\{\tau(X_i)\}. \tag{2.2}$$

The PATE is a function of the distribution of the potential outcomes in a population, whereas the SATE is a function of the potential outcomes themselves. The subtle distinction in their definitions leads to important differences in inferential and computational strategies, as will be discussed later. Traditionally, the SATE is of interest in randomized experiments where the target population is the specific sample, whereas the PATE is of interest in observational studies where the target population is the population from which the sample is drawn. In general, the choice of a causal estimand is determined by the scientific question in hand rather than statistical considerations. Note that although both the ITE and CATE are important in characterizing treatment effect heterogeneity, they are obviously different; however, these two estimands are sometimes conflated in the literature.

The *fundamental problem of causal inference* [14] is that, for each unit only the potential outcome corresponding to the actual treatment, $Y_i^{\mathrm{obs}} \equiv Y_i = Y_i(Z_i)$, is observed or *factual*, and the other potential outcome, $Y_i^{\mathrm{mis}} = Y_i(1 - Z_i)$, is missing or *counterfactual*. Therefore, additional assumptions are necessary to identify the causal effects. The key identifying assumptions concern the assignment mechanism, i.e. the process that determines which units get what treatment and hence which potential outcomes are observed or missing [15]. The vast majority of causal studies assume certain versions of an *ignorable assignment mechanism*, where the treatment assignment is independent of the potential outcomes conditional on some observed variables. Specifically, in the simple setting of a binary treatment at one time, ignorability consists of two sub-assumptions [15,16].

Assumption 2.1. ***(Ignorability)***. ***(a) Unconfoundedness***. $\Pr\{Z_i \mid Y_i(0), Y_i(1), X_i\} = \Pr(Z_i \mid X_i)$, *or equivalently* $Z_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\} \mid X_i$. ***(b) Overlap***. $0 < e(X_i) < 1$ *for all* i, *where* $e(x) \equiv \Pr(Z_i = 1 \mid X_i = x)$ *is the propensity score [16].*

The unconfoundedness assumption states that there is no unmeasured confounding, and the overlap assumption states that each unit has non-zero probability of being assigned to each treatment condition. These two assumptions together ensure that the conditional distribution of the potential outcomes is identifiable from observed data as

$$\mu_z(x) \equiv \mathbb{E}\{Y_i(z) \mid X_i = x\} = \mathbb{E}(Y_i \mid Z_i = z, X_i = x), \quad \text{for all } z, x. \tag{2.3}$$

Therefore, the CATE is identified as $\tau(x) = \mu_1(x) - \mu_0(x)$, and the PATE is identified as $\tau^{\mathrm{P}} = \mathbb{E}\{\mu_1(X_i) - \mu_0(X_i)\}$. This underlines the estimation strategy of *outcome modelling*: we can specify a model for the outcome function $\mu_z(x)$, and estimate the CATE by $\widehat{\tau}(x) = \widehat{\mu}_1(x) - \widehat{\mu}_0(x)$, and the

PATE by $\widehat{\tau}^{\text{reg}} = N^{-1}\sum_{i=1}^{N}\{\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\}$, where $\widehat{\mu}_z(x)$ is the estimated outcome model from the observed data.

In randomized experiments, the treatment assignment is known and controlled by the experimenters, and ignorability holds by design. In observational studies, the treatment assignment is unknown and uncontrolled, and ignorability at best holds approximately. A key concept in causal inference is *overlap and balance*, which refers to the similarity in the distributions of the covariates between the comparison groups. In general, as the two groups become more balanced, the causal estimates become less sensitive to the estimate strategy and model specification. In the ideal case of a randomized experiment, all—measured and unmeasured—covariates are balanced in expectation, i.e. they have the same multivariate distribution in the two treatment arms. Consequently, the simple difference-in-means estimator, $\widehat{\tau} = \{\sum_{i=1}^{N} Y_i Z_i/\sum_{i=1}^{N} Z_i\} - \{\sum_{i=1}^{N} Y_i(1-Z_i)/\sum_{i=1}^{N}(1-Z_i)\}$, is unbiased for τ^{S} and τ^{P}; furthermore, even a misspecified linear outcome model leads to a consistent estimate of τ^{S} [17]. On the contrary, in observational studies, the two groups are often imbalanced in many covariates, e.g. patients receiving the treatment may be generally sicker and older than those receiving the control. In such cases, directly comparing the difference in the outcomes between the two groups would give biased estimates of the causal effects. Moreover, the fit of an outcome model would rely on extrapolation in the regions where the two groups are poorly overlapped, and consequently outcome-model-based estimators, such as $\widehat{\tau}^{\text{reg}}$, are sensitive to the model specification. Therefore, a main effort in causal inference with observational data is to ensure overlap and balance to mimic a randomized experiment as closely as possible. This process does not involve the outcome and is referred to as the *design* stage, in contrast to the *analysis* stage, which utilizes the outcome and estimates causal effects given the design stage [18]. A causal analysis of an observational study usually has both design and analysis stages, in parallel with those of a randomized controlled experiment.

The propensity score plays a central role in causal inference with observational data, owing to its two special properties [16]. First, the propensity score is a balancing score in the sense that $Z_i \perp\!\!\!\perp e(X_i) \mid X_i$. This means that balancing the scalar propensity score balances the multivariate distribution of the covariates. Second, if a treatment assignment is unconfounded given X_i, then it is unconfounded given $e(X_i)$, that is, $Z_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\} \mid X_i$ implies $Z_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\} \mid e(X_i)$. In observational studies, $e(X_i)$ is usually unknown and needs to be estimated, e.g. via a logistic regression model of the treatment on the covariates.

The propensity score is usually used with matching, weighting, or stratification to achieve balance and estimate causal effects. Specifically, matching methods use a certain algorithm to find pairs of units in the two groups with similar covariates according to a distance metric, e.g. the propensity score or the Mahalanobis distance, and then calculate the difference in the average observed outcome between the groups in the matched sample [19–21]. Weighting methods assign a weight to each unit, so that the weighted distribution of the covariates in the two groups are balanced [22], and then calculate the weighted difference in the outcomes between the two groups. An important weighting scheme is inverse probability weighting (IPW), based on the identification formula of the PATE: $\tau^{\text{P}} = \mathbb{E}\{Z_i Y_i/e(X_i) - (1-Z_i)Y_i/(1-e(X_i))\}$. A corresponding IPW estimator [23] is $\widehat{\tau}^{\text{ipw}} = N^{-1}\sum_{i=1}^{N}\{Z_i Y_i/\widehat{e}(X_i) - (1-Z_i)Y_i/(1-\widehat{e}(X_i))\}$, where $\widehat{e}(X_i)$ denotes the estimated propensity score for unit i. One can further augment the IPW estimator by an outcome model to obtain a semiparametric efficient estimator [24]: $\widehat{\tau}^{\text{dr}} = \widehat{\tau}^{\text{reg}} + N^{-1}\sum_{i=1}^{N}\{Z_i R_i/\widehat{e}(X_i) - (1-Z_i)R_i/(1-\widehat{e}(X_i))\}$, where $R_i = Y_i - \widehat{\mu}_{Z_i}(X_i)$ is the residual from the outcome model. The IPW estimator $\widehat{\tau}^{\text{ipw}}$ is consistent for τ if the propensity score model is correct, and the outcome-model estimator $\widehat{\tau}^{\text{reg}}$ is consistent if the outcome model is correct. Because the bias of the estimator $\widehat{\tau}^{\text{dr}}$ is a product of the residual of the propensity score model and that of the outcome model, $\widehat{\tau}^{\text{dr}}$ is *doubly robust* in the sense that it is consistent if either the propensity score or the outcome model, but not necessarily both, is correctly specified [25]. Despite the seemingly different construction, matching estimators, with proper mathematical formulations, can be viewed as non-parametric versions of $\widehat{\tau}^{\text{ipw}}$, $\widehat{\tau}^{\text{reg}}$ and $\widehat{\tau}^{\text{dr}}$ based on nearest-neighbour regressions [26]. These are the main Frequentist estimation strategies for τ^{P} under ignorability.

When the target estimand is the CATE, the primary estimation strategy is outcome modelling. We will discuss how to specify the outcome model in §4(a).

3. General structure of Bayesian causal inference

(a) Basic factorization and versions of causal estimands

Because of the unavoidable missing potential outcomes, causal inference under the potential outcomes framework is inherently a missing data problem [6,15]. The Bayesian paradigm offers a unified framework for statistical inference with missing data and thus for causal inference [27]. Below we review the general structure of Bayesian causal inference that was first outlined in [15].

Four quantities are associated with each unit i, $\{Y_i(0), Y_i(1), Z_i, X_i\}$, where $\{Z_i, X_i, Y_i(Z_i)\}$ are observed but $Y_i(1-Z_i)$ is missing. Bayesian inference views all these quantities as random variables and centres around specifying a model for them. Based on the Bayesian model, we can draw inference on causal estimands—functions of the model parameters, covariates and potential outcomes—from the posterior predictive distributions of the parameters and the unobserved potential outcomes. Specifically, we assume the joint distribution of these random variables of all units is governed by a parameter $\theta = (\theta_X, \theta_Z, \theta_Y)$, conditional on which the random variables for each unit are *i.i.d.*. Then we can factorize the joint density $\Pr\{Y_i(0), Y_i(1), Z_i, X_i \mid \theta\}$ for each unit i as

$$\Pr\{Z_i \mid Y_i(0), Y_i(1), X_i; \theta_Z\} \cdot \Pr\{Y_i(0), Y_i(1) \mid X_i; \theta_Y\} \cdot \Pr(X_i; \theta_X). \tag{3.1}$$

The three terms in (3.1) represent the model for the assignment mechanism, potential outcomes, and covariates, respectively. Under ignorability, the assignment mechanism further reduces to the propensity score model $\Pr(Z_i \mid X_i; \theta_Z)$.

Before diving into the technical details, we first clarify the subtle but important difference between the Bayesian estimation of the PATE and SATE estimands. For the PATE, we rewrite the outcome-model-based identification formula in §2 as $\tau^{\mathrm{P}} = \int \{\mu_1(x;\theta_Y) - \mu_0(x;\theta_Y)\} F(\mathrm{d}x;\theta_X)$, which depends only on the unknown parameters θ_X and θ_Y. Therefore, Bayesian inference for the PATE requires obtaining posterior distributions of (θ_X, θ_Y). By contrast, the SATE τ^{S} is a function of the potential outcomes $\{Y_i(0), Y_i(1)\}_{i=1}^N$, which involves both observed and missing quantities. Bayesian inference for the SATE requires imputing the missing potential outcomes Y_i^{mis} from their posterior predictive distributions based on the outcome model, and consequently deriving the posterior distribution of τ^{S}.

However, in practice, we rarely model the possibly multi-dimensional covariates X_i, and instead condition on the observed values of the covariates. This is equivalent to replacing $F(x;\theta_X)$ with $\widehat{\mathbb{F}}_X$, the empirical distribution of the covariates. Therefore, most Bayesian causal inference (e.g. [9,28]) in fact focuses on the mixed average treatment effect (MATE) [6]

$$\tau^{\mathrm{M}} \equiv \int \tau(x;\theta_Y) \widehat{\mathbb{F}}_X(\mathrm{d}x) = N^{-1} \sum_{i=1}^{N} \tau(X_i;\theta_Y), \tag{3.2}$$

where $\tau(x;\theta_Y) = \tau(x)$ highlights the dependence on the parameter θ_Y. The MATE is a convenient approximation of the PATE and is particularly natural under the Bayesian paradigm. The difference between the MATE and SATE is subtle: the former equals the average of the CATE whereas the latter equals the average of the ITEs over the finite sample. Based on the posterior distributions, the PATE has the largest uncertainty, whereas the SATE has the smallest uncertainty. The distinction between these estimands is illustrated in the following example.

Example 3.1. [Covariate adjustment in a randomized experiment] Consider a completely randomized experiment with covariates X. Assume the true model for potential outcomes is

$$\begin{pmatrix} Y_i(1) \\ Y_i(0) \end{pmatrix} \mid (X_i, \beta_1, \beta_0, \sigma_1^2, \sigma_0^2, \rho) \sim \mathcal{N}\left(\begin{pmatrix} \beta_1' X_i \\ \beta_0' X_i \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_0 \\ \rho\sigma_1\sigma_0 & \sigma_0^2 \end{pmatrix} \right), \quad i = 1, \ldots, N.$$

This model implies two univariate normal marginal models: $Y_i(z) \mid X_i, \beta_z, \sigma_z^2 \sim \mathcal{N}(\beta_z' X_i, \sigma_z^2)$ for $z = 0, 1$. In this example, the CATE is $\tau(x) = (\beta_1 - \beta_0)'x$; the PATE, SATE and MATE are

$$\left.\begin{aligned} \tau^{\mathrm{P}} &= (\beta_1 - \beta_0)'\mathbb{E}(X_i), \\ \tau^{\mathrm{S}} &= N^{-1}\sum_{i=1}^{N}\{Y_i(1) - Y_i(0)\}, \\ \tau^{\mathrm{M}} &= (\beta_1 - \beta_0)'\bar{X}, \end{aligned}\right\} \tag{3.3}$$

respectively, where $\bar{X} = N^{-1}\sum_{i=1}^{N} X_i$ is the sample mean of the covariates.

(b) Posterior inference of causal effects

Regardless of the version of the target estimand, the following assumption is commonly adopted.

Assumption 3.2. ***(Prior independence)***. *The parameters for the models of assignment mechanism θ_Z, outcome θ_Y, and covariates θ_X are a priori distinct and independent.*

Assumption 3.2 imposes independent prior distributions for parameters $(\theta_X, \theta_Z, \theta_Y)$. It is unique to the Bayesian paradigm of causal inference. It is imposed primarily for modelling and computational convenience and may appear innocuous. However, as elaborated in §4(b), it may lead to unintended and undesirable implications in high-dimensional problems. Under assumptions 2.1 and 3.2, the joint posterior distribution of $\theta = (\theta_X, \theta_Z, \theta_Y)$ and the missing potential outcomes is proportional to

$$\Pr(\theta_X)\prod_{i=1}^{N}\Pr(X_i;\theta_X)\cdot\Pr(\theta_Z)\prod_{i=1}^{N}\Pr(Z_i \mid X_i;\theta_Z)\cdot\Pr(\theta_Y)\prod_{i=1}^{N}\Pr\{Y_i(1), Y_i(0) \mid X_i;\theta_Y\}. \tag{3.4}$$

From (3.4), the posterior distributions of θ_X and θ_Y, and consequently of τ^{P}, do not depend on the second component corresponding to the propensity score. Therefore, the propensity score model is *ignorable* in Bayesian inference of τ^{P}. The same ignorability argument applies to other estimands such as τ^{S}, τ^{M} and $\tau(x)$ [6,9,15]. Furthermore, inference of τ^{M} does not depend on the covariate model $\Pr(X_i;\theta_X)$. Because of this, it is essential to specify the outcome model $\Pr\{Y_i(1), Y_i(0) \mid X_i;\theta_Y\}$ in Bayesian causal inference.

By definition, $\tau^{\mathrm{P}} = \mathbb{E}\{Y_i(1)\} - \mathbb{E}\{Y_i(0)\}$ does not depend on the association between $Y_i(0)$ and $Y_i(1)$, denoted by the parameter ρ. Similarly, $\tau(x)$ does not depend on ρ, but τ^{S} does. So in the inference of τ^{P} and $\tau(x)$, we usually directly specify the marginal models $\Pr\{Y_i(z) \mid X_i;\theta_Y\}$ or equivalently $\Pr(Y_i \mid Z_i = z, X_i;\theta_Y)$ under ignorability [28]. The observed-data likelihood based on (3.4) becomes $\prod_{i:Z_i=1}\Pr(Y_i \mid Z_i = 1, X_i;\theta_Y)\prod_{i:Z_i=0}\Pr(Y_i \mid Z_i = 0, X_i;\theta_Y)$. Imposing a prior for θ_Y, we can proceed to infer θ_Y and subsequently τ^{P}, τ^{M}, or $\tau(x)$ using the usual Bayesian inferential procedures.

Bayesian inference of τ^{S} is more complex, because it depends on both $Y_i(0)$ and $Y_i(1)$ and thus requires posterior sampling of both θ_Y and $\mathbf{Y}^{\mathrm{mis}}$. The most common sampling strategy is through data augmentation: iteratively simulate θ and $\mathbf{Y}^{\mathrm{mis}}$ given each other and the observed data, namely from $\Pr(\theta_Y \mid \mathbf{Y}^{\mathrm{mis}}, \mathbf{Y}^{\mathrm{obs}}, \mathbf{Z}, \mathbf{X})$ and $\Pr(\mathbf{Y}^{\mathrm{mis}} \mid \mathbf{Y}^{\mathrm{obs}}, \mathbf{Z}, \mathbf{X};\theta_Y)$. The former, given the observed data and the imputed $\mathbf{Y}^{\mathrm{mis}}$, can be straightforwardly obtained by a complete-data analysis based on $\Pr\{\theta_Y \mid \mathbf{Y}(1), \mathbf{Y}(0), \mathbf{X}\} \propto \Pr(\theta_Y)\prod_{i=1}^{N}\Pr\{Y_i(1), Y_i(0) \mid X_i;\theta_Y\}$. The latter requires more elaboration. Specifically, we can show that $\Pr(\mathbf{Y}^{\mathrm{mis}} \mid \mathbf{Y}^{\mathrm{obs}}, \mathbf{Z}, \mathbf{X};\theta_Y)$ is proportional to $\prod_{i:Z_i=1}\Pr\{Y_i(0) \mid Y_i(1), X_i;\theta_Y\}\prod_{i:Z_i=0}\Pr\{Y_i(1) \mid Y_i(0), X_i;\theta_Y\}$. This renders that imputing the $\mathbf{Y}^{\mathrm{mis}}$ depends crucially on the joint distribution of $\{Y_i(1), Y_i(0)\}$. Because $Y_i(0)$ and $Y_i(1)$ are never jointly observed, the data provide no information about their association ρ. Unless the specific marginal model places constraints on ρ, the posterior distribution of ρ would be the same as its prior. Consequently, the posterior distribution of τ^{S} would be sensitive to the prior of ρ.

The above discussion prompts us to clarify the notion of identifiability in Bayesian inference. Under the Frequentist paradigm, a parameter is identifiable if any of its two distinct values

give two different distributions of the observed data. Under the Bayesian paradigm, there is no consensus. For example, Lindley [29] argued that all parameters are identifiable in Bayesian analysis because with proper prior distributions, posterior distributions are always proper. In this sense, ρ is identifiable. However, due to the fundamental problem of causal inference, there is no information in the data on ρ and it is reasonable to label it as non-identifiable. This is distinct from the parameters that the data provide direct information on, e.g. those in the marginal distributions of the outcomes in each arm, which are reasonable to label as identifiable. Lindley's perspective of identifiability blurs such distinction. A more informative perspective is provided by Gustafson [30], who argued that a parameter is *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, such as ρ. Another example of a partially identifiable parameter is $\Pr\{Y_i(1) > Y_i(0)\}$, which depends on ρ [6,31]. In this perspective, identifiability in Bayesian inference is no longer all-or-nothing; instead, it is a continuum in between. This issue motivates the strategy of *transparent parameterization*, where one separates identifiable and non-identifiable parameters, and treats the latter as sensitivity parameters in a sensitivity analysis [32–36]. More discussion will be given in §6(b).

Example 1 revisited We now illustrate the posterior inference of the causal estimands in example 3.1. Here, the parameters β's and σ's are identifiable, but ρ is not in the Frequentist sense. We fit a Bayesian linear regression model of Y_i on X_i to each observed arm z, with independent priors on (β_1, σ_1^2) and (β_0, σ_0^2). The observed likelihood factorizes into two parts: the data in treatment group $\{(X_i, Y_i) : Z_i = 1\}$ and control group $\{(X_i, Y_i) : Z_i = 0\}$ contribute to the likelihood of (β_1, σ_1^2) and (β_0, σ_0^2), respectively. For example, imposing the conventional conjugate normal-inverse χ^2 priors, we can draw from the posterior distribution of β and σ, and thus that of the MATE by plugging the posterior draws into the closed-form of τ^{M} in (3.3). To obtain the PATE, we would have to specify a multivariate model for $\Pr(X; \theta)$, and derive the posterior distribution of θ_X and then plug it into the closed form of τ^{P} in (3.3). This can also be implemented, e.g. via a Bayesian bootstrap step without a model, as described in the next paragraph. To obtain the SATE, we could specify a prior for ρ or fix it to a value. Given ρ and each draw of $(\beta_1, \beta_0, \sigma_1^2, \sigma_0^2)$, we can impute Y_i^{mis} as follows: for treated units, $Y_i^{\mathrm{mis}} = Y_i(0)$, and we draw $Y_i(0)$ from $\mathcal{N}(\beta_0' X_i + \rho\sigma_0/\sigma_1 \cdot (Y_i - \beta_1' X_i), \sigma_0^2(1-\rho^2))$; for control units, $Y_i^{\mathrm{mis}} = Y_i(1)$, and we draw $Y_i(1)$ from $\mathcal{N}(\beta_1' X_i + \rho\sigma_1/\sigma_0 \cdot (Y_i - \beta_0' X_i), \sigma_1^2(1-\rho^2))$. Plugging these posterior predictive draws of Y_i^{mis} and the observed outcomes into the definition of τ^{S}, we obtain its posterior distribution. We suggest varying the sensitivity parameter ρ from 0 to 1, which corresponds to conditionally independent potential outcomes and perfectly correlated potential outcomes, respectively.

An interesting alternative Bayesian strategy is through the Bayesian bootstrap [37], where the units are re-weighted with weights drawn from a Dirichlet distribution. The Bayesian bootstrap is a general strategy to simulate the posterior distribution of a parameter under a non-parametric model, which can be viewed as the limit of the inference under the Dirichlet Process prior [38]. This renders the Bayesian bootstrap relevant to causal inference in at least two ways. First, one can generate posterior samples from the distribution of $\Pr(X_i; \theta_X)$ without specifying a parametric model. This is desirable in inferring the population estimands like the PATE and the CATE [39]. However, how to integrate these samples of X into the inference of the target causal estimand is case-dependent and generally adds complexity to the analysis compared to the MATE. Second, the Bayesian bootstrap offers a general recipe for incorporating many standard Frequentist procedures into Bayesian inference. For example, Taddy *et al.* [40] used it to quantify the uncertainty in linear and tree-based methods for estimating the CATE. Chamberlain & Imbens [41] used it in M-estimation with an application to the setting of instrumental variables (see §7(a)). However, we view the Bayesian bootstrap approach as peripheral to Bayesian causal inference because it does not capitalize on arguably the main strength of Bayesian inference, namely, a unified inferential framework underpinned by the Bayes theorem with versatile choice of priors and outcome models.

4. Model specification

(a) Common specification of the outcome model

Section 3 shows that the central component in Bayesian inference of the CATE, PATE, and MATE is to specify the outcome model $\mu_z(x) = \mathbb{E}(Y_i \mid X_i = x, Z_i = z; \theta_Y)$. We can either model the two treatment groups jointly with a single function $\mu_z(x) = \mu(z, x)$ or model each group separately with two functions $\mu_1(x)$ and $\mu_0(x)$, known as S-learner or T-learner, respectively, in the literature [42]. The most basic outcome model is a linear regression: $\mu(z, x) = x + z + xz$, with Gaussian error terms, where the treatment-covariate interaction term xz captures the treatment effect heterogeneity. This model is equivalent to specifying a linear regression in each group. But the equivalence does not hold for nonlinear models; in fact, S-learners and T-learners with the same type of nonlinear models often lead to markedly different causal estimates.

Linear regressions are easy to implement and interpret, but they are often too restricted. In real-world problems, it is crucial to specify $\mu(z, x)$ flexibly enough to approximate the possibly complex underlying true data generating mechanism. This is particularly desirable as the recent focus in causal inference has been moving toward heterogeneous treatment effects. Outcome modelling is the most natural approach in these studies: one can simply specify an outcome model and derive the CATE as a function of the model parameters. There has been a rapidly increasing adoption of non-parametric and machine learning models for $\mu(z, x)$. One of the most widely used such models is based on regression trees. At a high level, regression trees partition the covariate space into non-overlapping regions and the prediction in each region is based solely on the data that fall in that region. The parameters of a regression tree characterize where to split the covariate space and how to predict the outcomes in a terminal node [43]. An ensemble of regression trees—usually referred to as forests—are often combined to improve the prediction.

Within the Bayesian paradigm, the Bayesian Additive Regression Tree (BART) [44] has become very popular for causal inference. BART places certain priors on the parameters of the regression trees to control the depth of the tree and the degree of shrinkage of the mean parameters in terminal nodes. Hill [9] first advocated using BART to specify the outcome model $\mu(z, x)$ in an S-learner. One can also specify a T-learner with a separate BART model for each treatment group $\mu_z(x)$. However, without any additional structure on the marginal models $\mu_z(x)$, T-learners often result in large variance of the treatment effects. Hahn *et al.* [8] proposed the *Bayesian Causal Forest* method based on an alternative reparametrization, $\mu(z, x) = g_1(x) + g_2(x)z$, where $g_1(x)$ models the distribution of $Y(0)$ and $g_2(x)$ represents the heterogeneous treatment effect, with a separate BART prior for $g_1(x)$ and $g_2(x)$. The BART models have a number of advantages, including fast computation, good performance of default choice of hyperparameters and available software. When a study has adequate covariate overlap, BART has been shown to outperform numerous competing methods, including (the Frequentist) random forests, in many empirical applications, e.g. [45,46]. Other Bayesian non-parametric models, such as Gaussian process [47], Dirichlet process [48–51], have also been considered for causal inference. We refer interested readers to [10] for a more detailed review of these methods.

(b) Challenges in high dimensions

Conducting statistical inference in high dimensions is challenging in general. We differentiate between two high-dimensional settings: (i) an outcome model with an infinite or a large number of parameters, regardless of the number of covariates, such as non-parametric and semiparametric models, and (ii) a large number of covariates. Both settings are increasingly common in causal inference, particularly in studies targeting the CATE. As discussed earlier, outcome modelling is the primary method in these settings, and Bayesian non-parametric models have become a mainstay of the model choice.

A straightforward application of the standard Bayesian non-parametric priors to outcome modelling is sometimes inadequate for causal inference, even with low dimensional covariates.

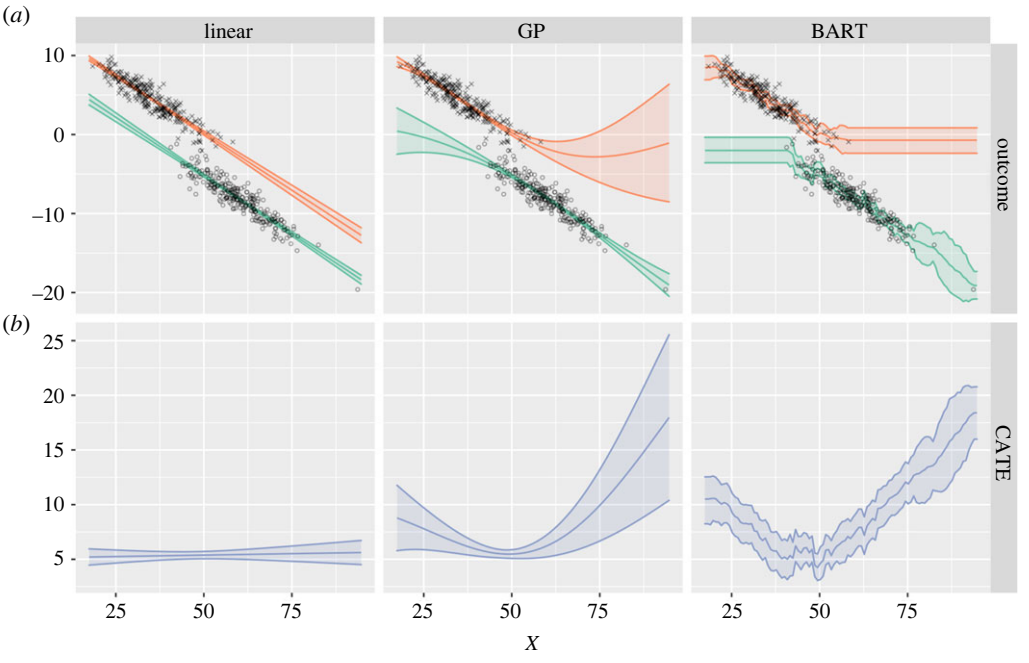

Figure 1. Example 4.1: estimates of means of the potential outcomes (*a*) and the CATE (*b*) and corresponding uncertainty band as a function of the single covariate by linear model, Gaussian Process, BART, respectively. Cross symbols: treated units; circles: control units. (Online version in colour.)

An important consideration is that a desirable prior should accurately reflect uncertainty according to the degree of covariate overlap because intuitively the uncertainty of causal estimates should increase as the degree of overlap decreases. Below, we reproduce a simple example in [52] (first constructed by Surya Tokdar) with a single covariate to illustrate this point.

Example 4.1. [Choice of priors in estimating the CATE] Consider a study with 250 treated and 250 control units. Each unit has a single covariate X that follows a Gamma distribution with mean 60 and 35 in the control ($Z_i = 0$) and treatment ($Z_i = 1$) group, respectively, and with s.d. 8 in both groups. To convey the main message, we consider a true outcome model with constant treatment effects: $Y_i(z) = 10 + 5z - 0.3X_i + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 1)$, where the CATE $\tau(x) = 5$ for all x. The scatterplots in the upper panel of figure 1 show that covariate overlap is good between the groups in the middle of the range of X (around 40 to 50), but deteriorates towards the tails of X.

To estimate the CATE, we fit an outcome model separately in each group: $\mu(z, x) = f_z(x) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. We choose three priors for $f_z(x)$: (i) a linear model $f_z(x) = \alpha_z + \beta_z x$ with a Gaussian prior for the coefficients; (ii) a BART prior similar to [9]; (iii) a Gaussian Process prior [53] with the covariance function specified by a Gaussian kernel with signal-to-noise ratio parameter ρ and inverse-bandwidth parameter λ: $(f_z(x_1), f_z(x_2), \ldots, f_z(x_N))' \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{ij} = \delta^2 \rho^2 \exp\{-\lambda^2 ||x_i - x_j||^2\}$). Figure 1 shows the posterior means of $\mu_z(X)$ and the CATE, with corresponding uncertainty band as a function of X. Here, we focus on the uncertainty quantification. In the region of good overlap, all three models lead to similar points and credible interval estimates of the CATE, but a marked difference emerges in the region of poor overlap. The linear model appears overconfident in estimating the CATE. The Gaussian process trades potential bias with wider credible interval as overlap decreases and produces a more adaptive uncertainty quantification. BART produces shorter error bars than the Gaussian Process (but wider than linear models), but the width of the credible interval remains similar regardless of the degree of overlap and thus is over confident in the presence of poor overlap.

Example 4.1 illustrates that, even with low dimensional covariates, standard Bayesian priors can have markedly different operating characteristics when the two groups are poorly overlapped, and not all priors can adaptively capture the uncertainty according to the degree of overlap. A primary reason for BART potentially underestimating uncertainty in poor overlap is its lack of smoothness, in contrast to the Gaussian process. Nonetheless, such a problem can be mitigated by soft decision trees as in [54].

When there are a large number of covariates, the Bayesian paradigm usually achieves regularization through sparsity-inducing priors for the outcome model, such as the spike-and-slab prior [55], the Bayesian LASSO [56], as well as the model averaging techniques [57–59]. The use of these methods in causal inference is surveyed in [10,50]. High-dimensional covariates pose additional complications to Bayesian causal inference. Robins & Ritov [60] pointed out that non-parametric estimates often have slow convergence rates in this regime, which translates into poor finite-sample performance. A central challenge in causal inference is that covariate overlap is rapidly diminishing as the covariates dimension increases, violating the overlap assumption that underpins standard causal analysis [61]. Lack of overlap exacerbates the usual inferential challenges—such as sparsity and slow convergence—in high-dimensional analysis. Even if we assume linear outcome models, we must carefully specify the priors on the regression coefficients. For example, Hahn *et al.* [7] showed that standard Bayesian regularization on the nuisance parameters may indirectly regularize important causal parameters and thus induce bias, namely the *regularization induced confounding*. This issue was rigorously investigated in [62]. Specifically, Linero [62] defines the selection bias as $\delta_z = \mathbb{E}(Y_i \mid Z_i = z) - \mathbb{E}\{Y_i(z)\}$, and showed that, under the seemingly innocuous prior independence assumption 3.2, many Bayesian regularization priors would *a priori* induce the selection bias δ_z to sharply concentrate around zero as the number of covariates, p, increases, to the extent that no amount of data would overcome such a bias. This implies that assumption 3.2 effectively acts as a strongly informative prior as p increases. Such a phenomenon is referred to as *prior dogmaticism* and is the Bayesian analogue of the aforementioned problem in Ritov *et al.* [63]. This line of research highlighted the importance of incorporating the propensity score in Bayesian causal inference [7,62,64,65], which echos the insights from the Frequentist *double machine learning* method [66,67]. Specifically, the regularized propensity score model or outcome model alone would not be sufficient for valid causal inference, but combining the two would achieve desirable convergence rate and finite sample performance in high-dimensional causal analysis.

5. The role of the propensity score

A major debate in Bayesian causal inference is the role of the propensity score, which characterizes the assignment mechanism. On the one hand, as shown in §3, under assumptions 2.1 and 3.2, the propensity score drops out from the likelihood and thus its value appears to be irrelevant in Bayesian causal inference, which seemingly only involves the outcome model and thus the analysis stage. On the other hand, §2 shows that the propensity score is ubiquitous in the Frequentist approach to causal inference, e.g. in constructing weighting, matching and doubly-robust estimators. Regardless of the mode of inference, the propensity score is essential in ensuring overlap and balance in the design stage of an observational study, which consequently reduces the sensitivity to the outcome model specification. Such sensitivity reduction is key to robust Bayesian causal inference, which is primarily based on outcome modelling. The literature has recognized the importance of incorporating the propensity scores into Bayesian causal inference, either in the design or the analysis stage, but there is no consensus on how to proceed. Below we review three existing strategies.

(a) Include the propensity score as a covariate in the outcome model

The propensity score was first proposed to be included as the *only* covariate in a Bayesian outcome model under ignorability, which would reduce the model complexity [68]. However,

as later pointed out by Zigler [59]: $\Pr\{Y(z) \mid X, e(X)\} = \Pr\{Y(z) \mid X\} \neq \Pr\{Y(z) \mid e(X)\}$. So using the propensity score as the single covariate in the outcome model would not lead to the target outcome distribution $\Pr\{Y(z) \mid X\}$, but using it as an additional covariate, i.e. specifying a model $\mu(z, x) = \mu(z, x, e(x))$, would. This specification is effectively conducting an outcome regression at each value of the propensity score, and thus can be viewed as a smoothed version of combining propensity score stratification and outcome modelling. In a sense, this specification provides a Bayesian analogue of the double robustness [52,69]. On the one hand, when the outcome model is correctly specified, $\mu(Z_i, X_i, e(X_i))$ reduces to $\mu(Z_i, X_i)$ because $e(X_i)$ is a function of X_i and thus is redundant regardless of its specification. On the other hand, when the outcome model is misspecified but the propensity model is correctly specified, the results are robust to the outcome model specification because the treatment and control groups are approximately balanced in covariates within each stratum of the propensity score. Various reparametrizations have been proposed. One example is to specify $\mu(z, x, e(x)) = g_1(x, e(x)) + g_2(z, x)$, with $g_1(\cdot)$ being a non-parametric model and $g_2(\cdot)$ being a parametric model. Little [70] adopted a penalized spline model of $e(x)$ for $g_1(\cdot)$. In the aforementioned *Bayesian Causal Forest*, Hahn *et al.* [8] imposed a separate BART model for $g_1(\cdot)$ and $g_2(\cdot)$, and demonstrated that adding the propensity score as an additional predictor in g_1 significantly improves the empirical estimation of the CATE.

This strategy is usually implemented in two stages: first estimate the propensity score as $\widehat{e}(X)$ and then plug it into the Bayesian outcome model $\mu(Z, X, \widehat{e}(X))$. Such a two-stage procedure is not dogmatically Bayesian, which generically refers to the procedure of specifying a model with parameters and prior distributions of these parameters and then use the Bayes theorem to obtain the posterior distributions of the parameters. A direct consequence is that this procedure may not properly propagate the uncertainty of estimating the propensity score in the outcome model [69]. A dogmatic Bayesian approach would jointly model $e(X; \theta_Z)$ and $\mu(Z, X, e(X); \theta_Y)$ and draw posterior inference of θ_Z and θ_Y simultaneously [71]. However, when the outcome model is misspecified, the joint-modelling approach would introduce a *feedback* problem, that is, the fit of the outcome model would inform the estimation of the propensity scores. This violates the unconfoundedness assumption, distorts the balancing property of the propensity score, and consequently leads to biased estimate of causal effects. A suggested remedy is to first fit a Bayesian model for e and then plug the posterior predictive draws of e into the outcome [11]. Such a two-stage procedure is still not dogmatically Bayesian, but provides more robust posterior inference to model misspecification empirically.

However, adding the propensity score into the outcome model is controversial conceptually, because the outcome model reflects the nature of the generating process of the potential outcomes, which arguably should not depend on how the treatment is assigned [72].

(b) Dependent priors

The Bayesian causal inference outlined in §3 rests on the prior independence assumption 3.2, without which the propensity score model cannot be ignored from the likelihood. But this assumption is not always plausible in real applications. Various priors that do not rely on this assumption have been proposed [47,63–65]. Below we show two simple examples.

The first example is due to [58] and is designed for simultaneous variable selection for the propensity score and outcome models. Specifically, assume a logistic propensity score model $\text{logit}\{\Pr(Z_i = 1 \mid X_i)\} = \alpha' X_i$ and a linear outcome model $Y_i \mid Z_i, X_i \sim \mathcal{N}(\tau Z_i + \beta' X_i, \sigma^2)$. Assume each of the jth components of the coefficients, α_j, follow the spike-and-slab prior [73]: $\alpha_j \mid \gamma_j^\alpha \sim (1 - \gamma_j^\alpha) I_0 + \gamma_j^\alpha \mathcal{N}(0, \sigma_\alpha^2)$, where γ_j^α is a latent indicator of whether X_j is included in the model and I_0 denotes the point mass at 0. A similar spike-and-slab prior is assumed for the coefficients of the outcome model with a latent inclusion indicator γ_j^β: $\beta_j \mid \gamma_j^\beta \sim (1 - \gamma_j^\beta) I_0 + \gamma_j^\beta \mathcal{N}(0, \sigma_\beta^2)$. Then assume the probability of the events $\{\gamma_j^\alpha = 0\}$ and $\{\gamma_j^\beta = 0\}$ are dependent *a priori*: $\Pr(\gamma_j^\beta = 1 \mid \gamma_j^\alpha = 1) / \Pr(\gamma_j^\beta = 0 \mid \gamma_j^\alpha = 1) = \omega$, where $\omega \in [1, \infty)$ is a dependence hyperparameter that controls the prior odds of including X_j into the outcome model when it is included in

the propensity score model. Larger ω implies stronger prior dependence between the variable selection in the two models.

The second example is due to [74]. Assume $Y_i(1) \mid X_i \sim \mathcal{N}(\mu_1, \sigma_1^2 e(X_i))$ and $Y_i(0) \mid X_i \sim \mathcal{N}(\mu_0, \sigma_0^2(1-e(X_i)))$, with flat priors on μ_1 and μ_0. If the propensity scores are known, the posterior mean of the PATE equals the Hajék version of the IPW estimator:

$$\widehat{\tau}^{\text{hajek}} = \frac{\sum_{i=1}^N Z_i Y_i / e(X_i)}{\sum_{i=1}^N Z_i / e(X_i)} - \frac{\sum_{i=1}^N (1-Z_i) Y_i / (1-e(X_i))}{\sum_{i=1}^N (1-Z_i)/(1-e(X_i))}.$$

If the propensity scores are unknown, then the posterior mean of the PATE is closely related to $\widehat{\tau}^{\text{hajek}}$ averaged over the posterior predictive distribution of the propensity scores. This strategy simply includes the propensity scores in the outcome model, but somewhat unusually in the conditional variances, rather than the conditional means of the potential outcomes.

Carefully designed dependent priors often achieve desirable finite sample results and are more reasonable in real world studies. However, specification of such priors is case-dependent, and there is no general solution.

(c) Posterior predictive inference

A general, albeit not dogmatically Bayesian, strategy is to specify both a propensity score model $e(X_i; \theta_Z)$ and an outcome model $\{\mu_1(X_i; \theta_Y), \mu_0(X_i; \theta_Y)\}$, and obtain posterior draws of $e(X_i; \theta_Z)$ and $\{\mu_1(X_i; \theta_Y), \mu_0(X_i; \theta_Y)\}$ from their respective posterior predictive distributions, and then plug these posterior draws into the doubly-robust estimator $\widehat{\tau}^{\text{dr}}$ [66,75]. A variance estimator of the resulting estimator $\widehat{\tau}^{\text{dr}}$ is given in [66]. In the same vein, Ding & Guo [76] incorporated the propensity score in Bayesian posterior predictive p-value. For the model with the Fisher's sharp null hypothesis of no causal effect for any units whatsoever (i.e. $Y_i(1) = Y_i(0)$ for all i), the procedure in [76] is equivalent to the Fisher randomization test averaged over the posterior predictive distribution of the propensity score. Simulations in [76] show the advantages of the Bayesian p-value compared to the Frequentist analogue. This perspective offers a straightforward and flexible strategy to integrate Bayesian modelling and common Frequentist procedures for causal inference and enables proper uncertainty quantification.

Besides the above three strategies, another general approach is through the aforementioned Bayesian bootstrap, which can be used to simulate the posterior distribution of any parameter that can be formulated as M-estimation or estimating equation [41,77]. As special cases, because the IPW estimator $\widehat{\tau}^{\text{ipw}}$ and the doubly robust estimator $\widehat{\tau}^{\text{dr}}$—both involving the propensity scores—are both solutions to estimating equations, they can be naturally combined with the Bayesian bootstrap to devise a Bayesian version. However, such an approach may be guilty of 'Bayesian for the sake of being Bayesian', and their methodological and practical value compared to competing methods is unclear.

6. Sensitivity analysis in observational studies

Unconfoundedness is a central assumption for causal inference. It holds by design in randomized experiments. However, its validity is fundamentally untestable in observational studies. Therefore, it is of great importance to assess the sensitivity of the results with respect to unmeasured confounding in any observational study. Such procedures are broadly called *sensitivity analysis*. Different classes of sensitivity analysis methods are characterized by the specific parametrization of confounding. Below we review the two most popular classes.

(a) Parametrization involving distributions with unmeasured confounders

The first parametrization used for sensitivity analysis is motivated by the intuition that a hidden confounder may completely explain away the association between the treatment and the outcome even after adjusting for observed covariates. In a historic debate, Fisher [78] hypothesized that

the strong association between cigarette smoking and lung cancer might be due to a hidden genetic factor as their 'common cause' or confounder. Cornfield *et al.* [79] derived an inequality showing that to explain away the observed association, the association between the unmeasured confounder and cigarette smoking must be larger than or equal to the association between cigarette smoking and lung cancer. Their work helped to initiate the field of sensitivity analysis.

Let U denote an unmeasured confounder and assume that unconfoundedness holds conditional on (X, U): $Z \perp\!\!\!\perp \{Y(0), Y(1)\} \mid X, U$. The joint distribution of all variables factorizes into

$$\Pr\{Y(1), Y(0), Z, X, U\} = \Pr\{Y(1), Y(0) \mid X, U\} \cdot \Pr(Z \mid X, U) \cdot \Pr(U \mid X) \cdot \Pr(X). \tag{6.1}$$

Under the factorization (6.1), sensitivity analysis requires us to specify the models for $\Pr\{Y(1), Y(0) \mid X, U\}$, $\Pr(Z \mid X, U)$ and $\Pr(U \mid X)$. In the special case of a binary Z, a binary Y and a discrete X (which can be thought as a stratified propensity score), Rosenbaum & Rubin [80] assumed a logistic model for Y given (Z, X, U), a logistic model for Z given (X, U), and a Bernoulli distribution for U, and treated the logistic regression coefficients of U and the probability parameter of U as the sensitivity parameters. They integrated out U in the complete-data likelihood and obtained the maximum likelihood estimates of τ^{P} over a plausible range of values of the sensitivity parameters. This method has been extended to more general settings in the Frequentist fashion in [81,82]. The Bayesian analogue of [80] is straightforward and can leverage the data augmentation algorithm to impute U to simplify the computation. Dorie *et al.* [83] extended this method to impose a Bayesian semiparametric model with a BART component for $\Pr\{Y(1), Y(0) \mid X, U\}$ to allow for model flexibility.

As an extension of [79], Ding & VanderWeele [84] treated the treatment-confounder (Z and U) and outcome-confounder (Y and U) associations as two sensitivity parameters, and derived analytical thresholds for them in order to explain away the observed treatment-outcome (Z and Y) association. Based on that theory, VanderWeele & Ding [85] further simplified by assuming the two associations to be the same and called the resulting threshold the *E-value*, as a measure of robustness of the causal conclusions with respect to unmeasured confounding. The E-value framework is model-free because it avoids modelling assumptions with U; it also avoids repeating the analysis over a range of sensitivity parameters as in the competing methods and thus is simple to calculate.

(b) Parametrization involving distributions of potential outcomes

The second parametrization is motivated by an alternative mathematical expression of the unconfoundedness assumption: $\Pr\{Y(z) \mid Z = 1, X\} = \Pr\{Y(z) \mid Z = 0, X\}$ for $z = 0, 1$, representing the fact that the units in the two randomized arms are comparable in terms of potential outcomes. This class of sensitivity analysis is based on sensitivity parameters that directly represent the difference between the distributions $\Pr\{Y(z) \mid Z = 1, X\}$ and $\Pr\{Y(z) \mid Z = 0, X\}$ instead of modelling the difference with an unobserved U. This is implemented in the context of time-varying treatments (see §7(b)) and Frequentist semiparametric estimation [86]. Franks *et al.* [34] pointed out the importance of distinguishing between model fit and sensitivity to unconfoundedness: the former involves identifiable parameters (e.g. the parameters in the model of the marginal distributions of the outcome $\Pr\{Y_i(z) \mid Z_i = z, X\}$) whereas the latter involves unidentifiable parameters (e.g. the association between $Y_i(1)$ and $Y_i(0)$). The merit of this parametrization is apparent in this perspective because it separates identifiable and unidentifiable parameters. Franks *et al.* [34] proceeded under the Bayesian paradigm and used a copula—parameters of which are the sensitivity parameters—to connect the two identifiable marginal distributions.

A related branch of sensitivity analysis is Rosenbaum's bounds [87]. His original formulation takes the association between Z and the potential outcomes conditional on observed X, denoted by Γ, as the sole sensitivity parameter for quantifying unmeasured confounding. He has also made connections to the parametrization in §6(a) [88]. Starting with a matched sample to mimic a randomized matched-pairs experiment, one can then repeat the Fisher randomization test on the sharp null hypothesis of no treatment effect given a range of Γ values, and find the threshold

of Γ at which the p-value of the test changes from significant to insignificant. This approach was later generalized to derive the bounds of a given estimator under different Γ values. Grounded in Fisherian randomization inference, Rosenbaum's framework does not have a natural Bayesian analogue.

Besides the above two classes, there are numerous other approaches to sensitivity analysis based on alternative parameterization of unmeasured confounding. However, a common criticism of various approaches to sensitivity analysis is that, in order to assess the consequence of the untestable unconfoundedness assumption, one has to make even more untestable assumptions, e.g. specifying models involving U. Moreover, sensitivity analysis, after all, is a secondary analysis in causal studies, and thus simple implementation and intuitive interpretation is much desired. The considerations underpin the dominance of the E-value method over other methods in practice, particularly in medicine and public health.

7. Complex assignment mechanisms

So far, we have discussed the simplest causal setting of an ignorable treatment at one time point. The basic formulation can be extended to many more complex assignment mechanisms. There are also many popular quasi-experimental designs that rely on identification strategies alternative to ignorability, e.g. regression discontinuity designs, difference-in-differences, synthetic controls. These designs are widely used in socioeconomic applications. Due to the space limit, below we will only briefly review two important extensions and refer interested readers to [89] for a review of quasi-experimental designs and related econometric methods.

(a) From instrumental variable to principal stratification

Instrumental variable (IV) is one of the most important techniques for causal inference in economics and social sciences. IVs are used in settings where dependence of the assignment on the potential outcomes cannot plausibly be ruled out, even conditional on observed covariates. An IV is a variable that provides a source of *exogenous* (or unconfounded) variation that helps identify causal effects. IV methods are based on a set of assumptions alternative to ignorability. Specifically, an IV satisfies three conditions: (i) it occurs before a treatment; (ii) it is independent of the treatment-outcome confounding; and (iii) it affects the outcome only through its (non-zero) effects on the treatment assignment. Finding a valid IV is challenging in observational studies and many clever natural experiments have been identified [89]. Given a valid IV, one can extract the causal effects of the treatment on an outcome by a two-stage least-squares (2SLS) estimator: first, fit a linear regression of the treatment on the IV; second, fit a linear regression of the outcome on the fitted value of the treatment from the first stage, the coefficient of which is taken as the causal effect of the treatment on the outcome. Covariates can be added in both stages.

The IV method has been developed within the structural equation model framework (see [89] for a review), and the 2SLS IV estimator may not correspond to a causal effect within the potential outcomes framework except for a few special cases. In a landmark paper, Angrist *et al.* [90] connects IV to the potential outcomes framework in the setting of randomized experiments with binary treatment and all-or-nothing compliance, with the initial random assignment playing the role of an IV. But many questions remain on the connection between the IV method and the potential outcomes framework in more general settings. Below we will describe the special setting of Angrist *et al.* [90].

We introduce some new notation. For unit i, let Z_i be the randomly assigned treatment (1 for the treatment and 0 for the control), and W_i be the actual treatment status (1 for the treatment and 0 for the control). When $Z_i \neq W_i$, non-compliance arises. Because W_i occurs post-assignment, it has two potential values, $W_i(0)$ and $W_i(1)$, with $W_i = W_i(Z_i)$. As before, the outcome Y_i has two potential outcomes, $Y_i(0)$ and $Y_i(1)$. Based on their joint potential status of the actual treatment $U_i = (W_i(1), W_i(0))$, the units fall into four compliance types: compliers $U_i = (1,0) = \text{co}$, never-takers $U_i = (0,0) = \text{nt}$, always-takers $U_i = (1,1) = \text{at}$ and defiers $U_i = (0,1) = \text{df}$ [90]. A key

property of U_i is that it is not affected by the treatment assignment, and thus can be regarded as a pre-treatment characteristic. Therefore, comparisons of $Y_i(1)$ and $Y_i(0)$ within the stratum of U_i have standard subgroup causal interpretations: $\tau_u = \mathbb{E}\{Y_i(1) - Y_i(0) \mid U_i = u\}$, for $u = \text{nt}, \text{co}, \text{at}, \text{df}$; τ_u are later called *principal causal effects*. The conventional causal estimand in clinical trials is the intention-to-treat effect that ignores the compliance information, which is the weighted average of the four stratum-specific effects: $\mathbb{E}\{Y_i(1) - Y_i(0)\} = \sum_u \pi_u \tau_u$, where $\pi_u = \Pr(U_i = u)$ is the proportion of the stratum u. The intention-to-treat effect measures the effect of the assignment instead of the actual treatment.

Due to the fundamental problem of causal inference, individual compliance stratum U_i is not observed. So the principal causal effects are non-identifiable without additional assumptions. Besides randomization of Z_i, Angrist *et al.* [90] make two additional assumptions: (i) monotonicity: $W_i(1) \geq W_i(0)$, and (ii) exclusion restriction: $Y_i(1) = Y_i(0)$ whenever $W_i(1) = W_i(0)$. Monotonicity rules out defiers, and exclusion restriction imposes that the assignment has zero effects among never-takers and always-takers. Then the compiler average causal effect is identified by

$$\tau_{\text{co}} \equiv \mathbb{E}\{Y(1) - Y(0) \mid U = \text{co}\} = \frac{\mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0)}{\mathbb{E}(W \mid Z = 1) - \mathbb{E}(W \mid Z = 0)},$$

which is exactly the probability limit of the two-stage least-squares estimator [90]. Because under monotonicity, only compilers' actual treatments are affected by the assignment, τ_{co} can be interpreted as the effect of the treatment.

We now describe the Bayesian inference of the IV set-up, first outlined in [91]. Without additional assumptions, the observed cells of Z and W consist of a mixture of units from more than one stratum. For example, the units who are assigned to the treatment arm and took the treatment $(Z = 1, W = 1)$ can be either always-takers or compliers. One must disentangle the causal effects for different compliance types from observed data. Therefore, model-based inference here resembles that of a mixture model. In Bayesian analysis, it is natural to impute the missing label U_i under some model assumptions. Specifically, six quantities are now associated with each unit, $\{Y_i(1), Y_i(0), W_i(1), W_i(0), X_i, Z_i\}$, four of which, $\{Y_i^{\text{obs}} = Y_i = Y_i(Z_i), W_i^{\text{obs}} = W_i = W_i(Z_i), Z_i, X_i\}$, are observed and the remaining two, $\{Y_i^{\text{mis}} = Y_i(1 - Z_i), W_i^{\text{mis}} = W_i(1 - Z_i)\}$, are unobserved. Assume the joint distribution of these random variables of all units is governed by a parameter θ, conditional on which the random variables for each unit are *iid*. We assume unconfoundedness $\Pr\{Z_i = 1 \mid X_i, W_i(1), W_i(0), Y_i(1), Y_i(0)\} = \Pr(Z_i = 1 \mid X_i)$, and impose a prior distribution $\Pr(\theta)$. Then the joint posterior distribution of θ and the missing potential outcomes are proportional to the complete-data likelihood as follows:

$$\Pr(\theta) \prod_{i=1}^{N} \Pr\{Y_i(0), Y_i(1) \mid U_i, X_i; \theta_Y\} \Pr(U_i \mid X_i; \theta_U) \Pr(X_i \mid \theta_X). \tag{7.1}$$

Without covariates, posterior inference of τ_u is straightforward because it is a function of θ_Y (see example 7.1 below). With covariates, we can condition on them and focus on a MATE estimand $\tau_u^{\text{M}} = N^{-1} \sum_{i=1}^{N} \mathbb{E}\{Y_i(1) - Y_i(0) \mid U_i = u, X_i\}$. The formula (7.1) suggests that we need to specify two models for inferring τ_u^{M}: (i) the compliance type model, $\Pr(U_i \mid X_i; \theta_U)$, and (ii) the outcome model, $\Pr\{Y_i(0), Y_i(1) \mid U_i, X_i; \theta_Y\}$. For example, we can specify a multinomial logistic regression for U_i and a generalized linear model for Y_i [91,92].

Using the same arguments as in §3, to infer population and mixed estimands, we only need to specify two marginal outcome models for $Y_i(1)$ and $Y_i(0)$ instead of a joint model, and do not need to impute the missing potential outcomes Y^{mis}. But we do need to impute the latent U_i, or, equivalently, the missing intermediate variable W^{mis}. We can simulate the joint posterior distribution $\Pr(\theta, \mathbf{W}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{Z}, \mathbf{X})$ by iteratively imputing the missing $\mathbf{U}$ from $\Pr(\mathbf{W}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{Z}, \mathbf{X}, \theta)$ and updating the posterior distribution of θ from $\Pr(\theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{W}^{\text{mis}}, \mathbf{Z}, \mathbf{X})$.

Below we illustrate the Bayesian procedure via a simple example of the IV approach.

Example 7.1. [Randomized experiment with one-sided non-compliance] Consider a randomized experiment with a binary outcome, where control units have no access to the treatment, i.e.

$W_i(0) = 0$ for all units. Therefore, we only have two strata: $U_i = \text{co}$ with $W_i(1) = 1$ and $U_i = \text{nt}$ with $W_i(1) = 0$, respectively, with $\pi_{\text{co}} + \pi_{\text{nt}} = 1$. Assume $Y_i(z) \mid U_i = \text{co} \sim \text{Bern}(p_{\text{co},z})$ for $z = 0, 1$, and $Y_i(1) = Y_i(0) \mid U_i = \text{nt} \sim \text{Bern}(p_{\text{nt}})$. So $\tau_{\text{co}} = p_{\text{co},1} - p_{\text{co},0}$. For simplicity, we impose conjugate priors on the parameters: $\pi_{\text{co}}, p_{\text{co},1}, p_{\text{co},0}, p_{\text{nt}}$ are *iid* Beta(1/2, 1/2). To sample the posterior distributions, the key is to impute the missing U_i's given the observed data. If $Z_i = 1$, then $W_i = 1$ implies $U_i = \text{co}$ and $W_i = 0$ implies $U_i = \text{nt}$, respectively. If $Z_i = 0$, then $W_i = 0$ and U_i is latent. For units with $(Z_i = 0, W_i = 0)$, we can impute $U_i = \text{co}$ with probability

$$\frac{\pi_{\text{co}} \cdot p_{\text{co},0}^{Y_i}(1 - p_{\text{co},0})^{1-Y_i}}{\pi_{\text{co}} \cdot p_{\text{co},0}^{Y_i}(1 - p_{\text{co},0})^{1-Y_i} + \pi_{\text{nt}} \cdot p_{\text{nt}}^{Y_i}(1 - p_{\text{nt}})^{1-Y_i}}$$

and $U_i = \text{nt}$ with the rest probability. With the imputed U_i's, we can sample the parameters from standard Beta posteriors: (i) sample π_{co} from $\text{Beta}(1/2 + \sum_{i=1}^N \mathbf{1}(U_i = \text{co}), 1/2 + \sum_{i=1}^N \mathbf{1}(U_i = \text{nt}))$ and obtain $\pi_{\text{nt}} = 1 - \pi_{\text{co}}$, (ii) sample $p_{\text{co},1}$ from $\text{Beta}(1/2 + \sum_{i=1}^N Z_i \mathbf{1}(U_i = \text{co}) Y_i, 1/2 + \sum_{i=1}^N Z_i \mathbf{1}(U_i = \text{co})(1 - Y_i))$, (iii) sample $p_{\text{co},0}$ from $\text{Beta}(1/2 + \sum_{i=1}^N (1 - Z_i)\mathbf{1}(U_i = \text{co}) Y_i, 1/2 + \sum_{i=1}^N (1 - Z_i)\mathbf{1}(U_i = \text{co})(1 - Y_i))$, and (iv) sample p_{nt} from $\text{Beta}(1/2 + \sum_{i=1}^N \mathbf{1}(U_i = \text{nt}) Y_i, 1/2 + \sum_{i=1}^N \mathbf{1}(U_i = \text{nt})(1 - Y_i))$. We iterate until convergence and obtain the posterior distribution of $\tau_{\text{co}} = p_{\text{co},1} - p_{\text{co},0}$. Imbens & Rubin [91] provided more detailed discussions.

Frangakis & Rubin [93] generalized the IV approach to principal stratification, a unified framework for causal inference with post-treatment confounding. In the simplest scenario, a post-treatment confounded variable lies in the causal pathway between the treatment and the outcome; it cannot be adjusted in the same fashion as a pre-treatment covariate in causal inference. A principal stratification with respect to a post-treatment variable is the classification of units based on the joint potential values of the post-treatment variable, and the stratum-specific effects are called *principal causal effects*, of which τ_{co} is a special case. The post-treatment variable setting includes a wide range of examples. For instance, in the non-compliance setting, the 'treatment' is the randomized treatment assignment, the 'post-treatment' variable is the actual treatment received, and the compliance types are the principal strata [91,92,94,95]. Zeng *et al.* [96] connects principal stratification to the local IV method with a continuous IV and binary treatment [97]. Other examples include censoring due to death [98], surrogate endpoints [99,100], regression discontinuity designs [101], time-varying treatments [102], and many more. The choices of target strata and thus estimands, interpretations, and identifying assumptions depend on specific applications, details of which are omitted here.

(b) Time-varying treatment and confounding

In real-world situations, subjects often receive treatments sequentially at multiple time points, and the treatment assignment at each time is affected by both baseline and time-varying confounders as well as previous treatments [103,104]. Such settings are referred to as time-varying, or sequential, or longitudinal treatments.

Consider a study where treatments are assigned at T time points. Let Z_{it} denote the treatment at time t for unit i $(i = 1, \ldots, N; t = 1, \ldots, T)$. At baseline $(t = 0)$, each unit i has time-invariant covariates L_{i0} measured; then after the treatment assignment at time $t - 1$ and prior to the assignment at time t, a set of time-varying confounders $L_{i,t-1}$ are measured, which include the intermediate measurements of the final outcome and the covariates that are affected by the previous treatments. For example, in a cancer study, baseline covariates can include sex, age, race and time-varying confounders can include intermediate cancer progression and other clinical traits such as blood pressure measured prior to the next treatment. Denote the observed and hypothetical treatment sequence of length t by $\bar{Z}_{it} = (Z_{i1}, \ldots, Z_{it})$ and $\bar{z}_t = (z_1, \ldots, z_t)$, respectively, and the sequence of time-varying confounders by $\bar{L}_{it} = (L_{i0}, L_{i1}, \ldots, L_{it})$. For each $\bar{z}_T$, there is a potential outcome $Y(\bar{z}_T)$. The final observed outcome $Y_i = Y_i(\bar{Z}_{iT})$, corresponding to the entire observed treatment sequence, is measured after treatment assigned at T. A common causal

estimand is the marginal effect comparing two pre-specified treatment sequences, $\bar{z}, \bar{z}' \in \{0,1\}^T$: $\tau_{\bar{z}_T, \bar{z}'_T} = \mathbb{E}\{Y_i(\bar{z}_T) - Y_i(\bar{z}'_T)\}$. For simplicity, below we drop the subscript i.

The central question to causal inference with sequential treatments is the role of the time-varying confounders L_t in the assignment mechanism. These variables are affected by the previous treatments and also affect the future treatment assignment and outcome. Much of the literature assumes a *sequentially ignorable* assignment mechanism [103], that is, the treatment at each time is unconfounded conditional on the observed history, which consists of past treatments $\bar{Z}_{t-1}$ and time-varying confounders $\bar{L}_{t-1}$, as stated below.

Assumption 7.2. *(Sequential Ignorability).* $\Pr\{Z_t \mid \bar{Z}_{t-1}, \bar{L}_{t-1}, Y(\bar{z}_t) \text{ for all } \bar{z}_t\} = \Pr(Z_t \mid \bar{Z}_{t-1}, \bar{L}_{t-1})$ *for* $t = 1, \ldots, T$.

A full Bayesian approach to time-varying treatments [105] would specify a joint model for treatment assignment Z_t and time-varying confounders L_t at all time points as well as all the potential outcomes $Y(\bar{z}_T)$, and then derive the posterior predictive distributions of the missing potential outcomes and thus of the estimands. This procedure is a straightforward extension from the structure introduced in §3. However, the joint modelling approach is rarely used because it quickly becomes intractable as the time T and the number of time-varying confounders increases.

Instead, most of the Bayesian methods are grounded in the g-computation. Under sequential ignorability, the average potential outcome $\mathbb{E}\{Y_i(\bar{z}_T)\}$ is identified from the observed data via the g-formula [103]:

$$\mathbb{E}\{Y_i(\bar{z}_T)\} = \sum_{L_0, L_1, \ldots, L_{T-1}} \mathbb{E}(Y \mid \bar{Z}_T = \bar{z}_T, \bar{L}_{T-1}) \cdot \Pr(L_{T-1} \mid \bar{Z}_{T-1} = \bar{z}_{T-1}, \bar{L}_{T-2}) \\ \cdots \Pr(L_1 \mid Z_1 = z_1, L_0) \cdot \Pr(L_0). \tag{7.2}$$

To operationalize the g-formula, we can specify models for all the components of (7.2), including an outcome model $\Pr(Y \mid \bar{Z}_T = \bar{z}_T, \bar{L}_{T-1})$ and a model for the time-varying confounders L_t at each time t, $\Pr(L_t \mid \bar{Z}_t = \bar{z}_t, \bar{L}_{t-1})$. The g-formula is in essence an extension of the outcome regression approach to time-varying treatments. The Bayesian version of the g-computation would specify a Bayesian model for each component in the g-formula (7.2) and then combine the posterior draws of the parameters to obtain the posterior distribution of the estimands. Below we present an illustrative example of Bayesian g-computation due to [106].

Example 7.3. [Bayesian g-computation with two periods] Consider the simplest possible scenario with two time periods, binary covariates and a binary outcome. Let L_0 be a binary baseline covariate, Z_1 is a binary treatment at time 1, L_1 is a binary time-varying covariate between time 1 and 2, Z_2 is a binary treatment, and Y is a binary outcome. To obtain the posterior distribution of

$$\mathbb{E}\{Y(z_1, z_2)\} = \sum_{l_0=0,1} \sum_{l_1=0,1} \Pr(Y = 1 \mid Z_1 = z_1, Z_2 = z_2, L_0 = l_0, L_1 = l_1) \\ \cdot \Pr(L_1 = l_1 \mid Z_1 = z_1, L_0 = l_0) \cdot \Pr(L_0 = l_0), \tag{7.3}$$

it suffices to obtain the posterior distributions of the probabilities in (7.3). Assuming the standard Beta(1/2, 1/2) conjugate priors. We can obtain the posterior of the probabilities as follows: (i) sample $\Pr(Y = 1 \mid Z_1 = z_1, Z_2 = z_2, L_0 = l_0, L_1 = l_1)$ from $\text{Beta}(1/2 + \sum_{i=1}^N \mathbf{1}(Z_{i1} = z_1, Z_{i2} = z_2, L_{i0} = l_0, L_{i1} = l_1) Y_i, 1/2 + \sum_{i=1}^N \mathbf{1}(Z_{i1} = z_1, Z_{i2} = z_2, L_{i0} = l_0, L_{i1} = l_1)(1 - Y_i))$, (ii) sample $\Pr(L_1 = 1 \mid Z_1 = z_1, L_0 = l_0)$ from $\text{Beta}(1/2 + \sum_{i=1}^N \mathbf{1}(Z_{i1} = z_1, L_{i0} = l_0) L_{i1}, 1/2 + \sum_{i=1}^N \mathbf{1}(Z_{i1} = z_1, L_{i0} = l_0)(1 - L_{i1}))$, and (iii) sample $\Pr(L_0 = 1)$ from $\text{Beta}(1/2 + \sum_{i=1}^N L_{i0}, 1/2 + \sum_{i=1}^N (1 - L_{i0}))$. With these ingredients and (7.3), we can obtain the posterior distributions of $\mathbb{E}\{Y(z_1, z_2)\}$'s and their contrasts $\sum_{z_1, z_2} c(z_1, z_2) \mathbb{E}\{Y(z_1, z_2)\}$.

G-computation quickly becomes complex as the number of time periods T and time-varying confounders increases, which requires specifying a large number of models. Then it is necessary to impose more structural restrictions on the data-generating process. However, Robins &

Wasserman [107] showed that unsaturated models might rule out the null hypothesis of zero causal effect *a priori*, a phenomenon termed the '*g*-null paradox'.

A popular alternative strategy is the marginal structural model [104], which generalizes IPW to time-varying treatments. Saarela *et al.* [108] devised a Bayesian version of the marginal structural model via the Bayesian bootstrap. Because the marginal structural model relies on IPW, a key component in its implementation is to estimate the propensity scores and ensure overlap at each time point. However, overlap between different treatment paths usually becomes limited as the number of time periods increases, rendering the marginal structural model sensitive to extreme weights.

The above discussion focuses on *static* treatment sequences. Another important class of time-varying treatment is the *dynamic* treatment regime, which consists of a sequence of decision rules, one per time point of intervention, that determines how to individualize treatments to units based on evolving treatment and covariate history. Inferring optimal dynamic treatment regimes requires combining causal inference and decision theory techniques and is closely related to reinforcement learning. See [109] for a review. Due to the space limit, we omit the discussion of the closely related topics of Bayesian multi-armed bandit [110] and Bayesian reinforcement learning [111].

8. Discussion

This paper reviews the Bayesian approach to causal inference under the potential outcomes framework. We discussed the causal estimands, identification strategies, the general structure of Bayesian inference of causal effects, and sensitivity analysis. We highlight issues that are unique to Bayesian causal inference, including the role of the propensity score, definition of identifiability, the choice of priors in both low- and high-dimensional regimes. In particular, under ignorability and prior independence, the propensity score is seemingly irrelevant for the posterior distributions of the causal parameters. However, we pointed out that even in this setting, the propensity score and more generally the design stage plays a central role in obtaining robust Bayesian causal inference. Regardless of the mode of inference, a critical step in causal inference with observational data is to ensure adequate covariate overlap and balance in the design or analysis stages. In high dimensions, such a task is particularly challenging and what is the optimal practice remains an open question.

The Bayesian approach offers several advantages for causal inference. First and most importantly, by enabling imputation of all missing potential outcomes, the Bayesian paradigm provides a unified inferential framework for any causal estimand. This is particularly appealing for inferring complex estimands such as the conditional average treatment effects or individual treatment effects as well as partially identifiable causal estimands such as the principal strata causal effects. In contrast, the Frequentist approach to these problems needs to be customized for each scenario, and the inference usually relies on bounds or asymptotic arguments, which are often either non-informative or questionable in cases like individual treatment effects. Second, the automatic uncertainty quantification of any estimand renders it straightforward to combine causal inference and decision theory for dynamic decision-making, e.g. in personalized medicine. Third, the Bayeian approach naturally incorporates prior knowledge into a causal analysis, e.g. in evaluating spatially correlated treatments and/or outcomes. Fourth, there is rich collection of Bayesian models for complex data with limited Frequentist counterparts. A few examples are (i) spatial or temporal data, (ii) functional data, and (iii) interference, i.e. when the SUTVA assumption is violated. In these settings, special care must be taken on issues key to causal inference such as defining relevant estimands and ensuring overlap. Moreover, it is important to ensure that the Bayesian models are coherent to the model-free identification assumptions such as ignorability. For example, adding spatial random effects into an outcome model may inadvertently bias the coefficient of the treatment variable as the estimate of a causal effect [112]. Research on Bayesian analysis of these topics has been rapidly increasing [10,112–115] and is expected to continue to grow.

Despite the above advantages, the theory and practice in causal inference has long been dominated by non-Bayesian methods. One reason is that many popular Frequentist techniques, such as matching and weighting, as well as Fisherian randomization test, do not require specifying outcome models and prior distribution of parameters, and thus offer a perception of 'model-free' or 'objective'. This is appealing to many applied researchers. Another reason is that the Bayesian approach requires more advanced computing and programming, which may not be readily available to many practitioners. The **Stan** programming language [116] mitigates some of these issues, but Bayesian computation remains inaccessible to most domain scientists. To popularize Bayesian causal inference in practice, it is crucial to provide (i) more examples of successful Bayesian applications with clear advantages over other inferential modes, e.g. [45], (ii) accessible tutorials, ideally with generalizable computer code and illustrations of important scientific problems and (iii) developing and disseminating user-friendly, general purpose software packages.

We have occasionally commented on whether a method is dogmatically Bayesian in the discussion. However, we do not regard the conceptual purity of being dogmatically Bayesian, *per se*, as advantageous, nor should it be the motivating goal in real applications. When a quasi-Bayesian method outperforms its dogmatically Bayesian counterpart (if available) with methodological footing and empirical evidence, as the example of adding estimated propensity score in an outcome model in §5(a), we would endorse the former over the latter. We also doubt the value of devising a Bayesian version of an established Frequentist method without clear theoretical or practical advantages. As a general view, we believe whether to choose a Bayesian approach should be dictated by its practical utility in a specific context rather than an unconditional commitment to the Bayesian doctrine. For causal inference and perhaps everything in statistics, being Bayesian should be a tool, not a goal.

Data accessibility. This article has no additional data.

Authors' contributions. F.L.: conceptualization, formal analysis, investigation, methodology, writing—original draft, writing—review and editing; P.D.: formal analysis, investigation, methodology, writing—original draft, writing—review and editing; F.M.: methodology, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. P.D.'s research is partially funded by the US National Science Foundation grant no. 1945136. F.M. thanks the Department of Excellence 2018–2022 funding provided by the Italian Ministry of University and Research (MUR).

Acknowledgements. The authors are grateful to Joey Antonelli, Yuansi Chen, Sid Chib, Ruobin Gong, Guido Imbens, Zhichao Jiang, Antonio Linero, Georgia Papadogeorgou, Donald Rubin, Surya Tokdar, Mike West, Jason Xu, Cory Zigler, Anqi Zhao and two anonymous reviewers for discussions and suggestions.

References

1. Neyman J. 1923 On the application of probability theory to agricultural experiments: essay on principles, §9. Masters Thesis. Portions translated into english by D. Dabrowska and T. Speed (1990) in *Stat. Sci.*, pp. 465–472.
2. Rubin DB. 1974 Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Edu. Psychol.* **66**, 688–701. (doi:10.1037/h0037350)
3. Rubin DB. 1975 Bayesian inference for causality: the role of randomization. In *Proc. of Social Statistics Section of Am Stat. Assoc.*, pp. 233–239.
4. Pearl J. 2000 *Causality: models, reasoning, and inference*. New York, NY: Cambridge University Press.
5. Peters J, Bühlmann P, Meinshausen N. 2016 Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. B* **78**, 947–1012. (doi:10.1111/rssb.12167)
6. Ding P, Li F. 2018 Causal inference: a missing data perspective. *Stat. Sci.* **33**, 214–237. (doi:10.1214/18-STS645)

7. Hahn PR, Carvalho CM, Puelz F, He J. 2018 Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal.* **13**, 163–182. (doi:10.1214/16-BA1044)
8. Hahn PR, Murray JS, Carvalho CM. 2020 Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal.* **15**, 965–1056. (doi:10.1214/19-BA1195)
9. Hill JL. 2011 Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.* **20**, 217–240. (doi:10.1198/jcgs.2010.08162)
10. Linero AR, Antonelli JL. 2022 The how and why of Bayesian nonparametric causal inference. *Wiley Interdiscip. Rev.: Comput. Stat.* **15**, e1583.
11. Zigler CM, Dominici F. 2014 Uncertainty in propensity score estimation: Bayesian methods for variable selection and model averaged causal effects. *J. Am. Stat. Assoc.* **109**, 95–107. (doi:10.1080/01621459.2013.869498)
12. Dawid AP. 1979 Conditional independence in statistical theory. *J. R. Stat. Soc. B* **41**, 1–15.
13. Rubin DB. 1980 Comment on 'Randomization analysis of experimental data: the Fisher randomization test' by D. Basu. *J. Am. Stat. Assoc.* **75**, 591–593.
14. Holland PW. 1986 Statistics and causal inference. *J. Am. Stat. Assoc.* **81**, 945–960. (doi:10.1080/01621459.1986.10478354)
15. Rubin DB. 1978 Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* **6**, 34–58. (doi:10.1214/aos/1176344064)
16. Rosenbaum PR, Rubin DB. 1983 The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55. (doi:10.1093/biomet/70.1.41)
17. Lin W. 2013 Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *Ann. Appl. Stat.* **7**, 295–318. (doi:10.1214/12-AOAS583)
18. Rubin DB. 2007 The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat. Med.* **26**, 20–36. (doi:10.1002/sim.2739)
19. Abadie A, Imbens GW. 2011 Bias corrected matching estimators for average treatment effects. *J. Bus. Econom. Stat.* **29**, 1–11. (doi:10.1198/jbes.2009.07333)
20. Abadie A, Imbens GW. 2006 Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**, 235–267. (doi:10.1111/j.1468-0262.2006.00655.x)
21. Rubin DB. 2006 *Matched sampling for causal effects*. Cambridge, UK: Cambridge University Press.
22. Li F, Morgan KL, Zaslavsky AM. 2018 Balancing covariates via propensity score weighting. *J. Am. Stat. Assoc.* **113**, 390–400. (doi:10.1080/01621459.2016.1260466)
23. Rosenbaum PR. 1987 Model-based direct adjustment. *J. Am. Stat. Assoc.* **82**, 387–394. (doi:10.1080/01621459.1987.10478441)
24. Robins JM, Rotnitzky A, Zhao LP. 1994 Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.* **89**, 846–866. (doi:10.1080/01621459.1994.10476818)
25. Bang H, Robins JM. 2005 Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–972. (doi:10.1111/j.1541-0420.2005.00377.x)
26. Lin Z, Ding P, Han F. 2021 Estimation based on nearest neighbor matching: from density ratio to average treatment effect. Preprint (https://arxiv.org/abs/2112.13506).
27. Rubin DB. 1976 Inference and missing data. *Biometrika* **63**, 581–592. (doi:10.1093/biomet/63.3.581)
28. Chib S. 2007 Analysis of treatment response data without the joint distribution of potential outcomes. *J. Econom.* **140**, 401–412. (doi:10.1016/j.jeconom.2006.07.009)
29. Lindley DV. 1972 *Bayesian statistics: a review*. SIAM.
30. Gustafson P. 2015 *Bayesian inference for partially identified models: exploring the limits of limited data*. New York, NY: CRC Press.
31. Lu J, Ding P, Dasgupta T. 2018 Treatment effects on ordinal outcomes: causal estimands and sharp bounds. *J. Educ. Behav. Stat.* **43**, 540–567. (doi:10.3102/1076998618776435)
32. Daniels MJ, Hogan JW. 2008 *Missing data in longitudinal studies: strategies for Bayesian modeling and sensitivity analysis*. London, UK: Chapman and Hall/CRC.
33. Ding P, Dasgupta T. 2016 A potential tale of two-by-two tables from completely randomized experiments. *J. Am. Stat. Assoc.* **111**, 157–168. (doi:10.1080/01621459.2014.995796)

34. Franks AM, D'Amour A, Feller A. 2020 Flexible sensitivity analysis for observational studies without observable implications. *J. Am. Stat. Assoc.* **115**, 1730–1746. (doi:10.1080/01621459.2019.1604369)
35. Gustafson P. 2009 What are the limits of posterior distributions arising from nonidentified models, why should we care? *J. Am. Stat. Assoc.* **104**, 1682–1695. (doi:10.1198/jasa.2009.tm08603)
36. Richardson TS, Evans RJ, Robins JM. 2010 Transparent parameterizations of models for potential outcomes. In *Bayesian Statistics*, vol. 9 (eds JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, M West), pp. 569–610. Oxford, UK: Oxford University Press.
37. Rubin DB. 1981 The Bayesian bootstrap. *Ann. Stat.* **9**, 130–134. (doi:10.1214/aos/1176345338)
38. Ferguson TS. 1973 A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1**, 209–230. (doi:10.1214/aos/1176342360)
39. Oganisian A, Mitra N, Roy JA. 2020 Hierarchical Bayesian bootstrap for heterogeneous treatment effect estimation. (https://arxiv.org/abs/2009.10839).
40. Taddy M, Gardner M, Chen L, Draper D. 2016 A nonparametric Bayesian analysis of heterogenous treatment effects in digital experimentation. *J. Bus. Econ. Stat.* **34**, 661–672. (doi:10.1080/07350015.2016.1172013)
41. Chamberlain G, Imbens GW. 2014 Nonparametric applications of Bayesian inference. *J. Bus. Econ. Stat.* **21**, 12–18. (doi:10.1198/073500102288618711)
42. Künzel SR, Sekhon JS, Bickel PJ, Yu B. 2019 Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl Acad. Sci.* **116**, 4156–4165. (doi:10.1073/pnas.1804597116)
43. Breiman L, Friedman JH, Olshen RA, Stone CJ. 2017 *Classification and regression trees*. Routledge.
44. Chipman HA, George EI, McCulloch RE. 2010 BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4**, 266–298. (doi:10.1214/09-AOAS285)
45. Dorie V, Hill J, Shalit U, Scott M, Cervone D. 2019 Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *Stat. Sci.* **4**, 43–68. (doi:10.1214/18-STS667)
46. Hu L, Ji J, Li F. 2021 Estimating heterogeneous survival treatment effect in observational data using machine learning. *Stat. Med.* **40**, 4691–4713. (doi:10.1002/sim.9090)
47. Ray K, van der Vaart A. 2020 Semiparametric Bayesian causal inference. *Ann. Stat.* **48**, 2999–3020. (doi:10.1214/19-AOS1919)
48. Chib S, Hamilton BH, 2002 Semiparametric Bayes analysis of longitudinal data treatment models. *J. Econom.* **110**, 67–89. (doi:10.1016/S0304-4076(02)00122-7)
49. Karabatsos G, Walker SG. 2012 A Bayesian nonparametric causal model. *J. Stat. Plan. Inference* **142**, 925–934. (doi:10.1016/j.jspi.2011.10.013)
50. Oganisian A, Roy JA. 2021 A practical introduction to Bayesian estimation of causal effects: parametric and nonparametric approaches. *Stat. Med.* **40**, 518–551. (doi:10.1002/sim.8761)
51. Roy J, Lum KJ, Zeldow B, Dworkin JD, Re III VL, Daniels MJ. 2018 Bayesian nonparametric generative models for causal inference with missing at random covariates. *Biometrics* **74**, 1193–1202. (doi:10.1111/biom.12875)
52. Papadogeorgou G, Li F. 2020 Discussion for 'Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects'. *Bayesian Anal.* **15**, 1007–1013.
53. Williams CK, Rasmussen CE. 2006 *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.
54. Linero AR, Yang Y. 2018 Bayesian regression tree ensembles that adapt to smoothness and sparsity. *J. R. Stat. Soc. B* **80**, 1087–1110. (doi:10.1111/rssb.12293)
55. Antonelli JL, Parmigiani G, Dominici F. 2019 High-dimensional confounding adjustment using continuous spike and slab priors. *Bayesian Anal.* **14**, 805–828. (doi:10.1214/18-BA1131)
56. Park T, Casella G. 2008 The Bayesian lasso. *J. Am. Stat. Assoc.* **103**, 681–686. (doi:10.1198/016214508000000337)
57. Raftery AE, Madigan D, Hoeting JA. 1997 Bayesian model averaging for linear regression models. *J. Am. Stat. Assoc.* **92**, 179–191. (doi:10.1080/01621459.1997.10473615)

58. Wang C, Parmigiani G, Dominici F. 2012 Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* **68**, 661–671. (doi:10.1111/j.1541-0420.2011.01731.x)
59. Zigler CM. 2016 The central role of Bayes' theorem for joint estimation of causal effects and propensity scores. *Am. Stat.* **70**, 47–54. (doi:10.1080/00031305.2015.1111260)
60. Robins JM, Ritov Y. 1997 Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Stat. Med.* **16**, 285–319. (doi:10.1002/(SICI)1097-0258(19970215)16:3<285::AID-SIM535>3.0.CO;2-%23)
61. D'Amour A, Ding P, Feller A, Lei L, Sekhon J. 2021 Overlap in observational studies with high-dimensional covariates. *J. Econom.* **221**, 644–654.
62. Linero AR. 2021 In nonparametric and high-dimensional models, Bayesian ignorability is an informative prior. Preprint (https://arxiv.org/abs/2111.05137).
63. Ritov Y, Bickel BJ, Gamst AC, Kleijn BJ. 2014 The Bayesian analysis of complex, high-dimensional models: can it be CODA. *Stat. Sci.* **29**, 619–639. (doi:10.1214/14-STS483)
64. Harmeling S, Touissant M. 2007 Bayesian estimators for Robins-Titov's problem. Technical report, School of Informatics, University of Edinburgh. See https://argmin.lis.tu-berlin.de/papers/07-harmeling-tr.pdf.
65. Sims CA. 2012 Robins-wasserman, round *N*. See http://sims.princeton.edu/yftp/WassermanExmpl/WassermanR4a.pdf.
66. Antonelli JL, Papadogeorgou G, Dominici F. 2022 Causal inference in high dimensions: a marriage between Bayesian modeling and good frequentist properties. *Biometrics* **78**, 100–114. (doi:10.1111/biom.13417)
67. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W. 2017 Double/debiased/neyman machine learning of treatment effects. *Am. Econom. Rev.* **107**, 261–265. (doi:10.1257/aer.p20171038)
68. Rubin DB. 1985 The use of propensity score in applied Bayesian inference. In *Bayesian statistics*, vol. 2 (eds JM Bernardo, MH DeGroot, DV Lindley, AFM Smith), pp. 463–472. North-Holland: Elsevier Science Publisher B.V.
69. Zigler CM, Watts K, Yeh RW, Wang Y, Coull BA, Dominici F. 2013 Model feedback in Bayesian propensity score estimation. *Biometrics* **69**, 263–273. (doi:10.1111/j.1541-0420.2012.01830.x)
70. Little RJA, An H. 2004 Robust likelihood-based analysis of multivariate data with missing values. *Stat. Sin.* **14**, 949–968.
71. McCandless LC, Gustafson P, Austin PC. 2009 Bayesian propensity score analysis for observational data. *Stat. Med.* **28**, 94–112. (doi:10.1002/sim.3460)
72. Robins JM, Hernán MA, Wasserman L. 2015 Discussion of 'Bayesian estimation of marginal structural models' by Saarela *et al*. *Biometrics* **71**, 293–296. (doi:10.1111/biom.12273)
73. George EI, McCulloch RE. 1997 Approaches for Bayesian variable selection. *Stat. Sin.* **1**, 339–373.
74. Little RJA. 2004 To model or not to model? Competing modes of inference for finite population sampling. *J. Am. Stat. Assoc.* **99**, 546–556. (doi:10.1198/016214504000000467)
75. Saarela O, Belzile LR, Stephens DA. 2016 A Bayesian view of doubly robust causal inference. *Biometrika* **103**, 667–681. (doi:10.1093/biomet/asw025)
76. Ding P, Guo T. 2023 Posterior predictive propensity scores and *p*-values. *Observational Studies* **9**, 3–18. (doi:10.1353/obs.2023.0015)
77. Lyddon SP, Holmes CC, Walker SG. 2019 General Bayesian updating and the loss-likelihood bootstrap. *Biometrika* **106**, 465–478. (doi:10.1093/biomet/asz006)
78. Ferguson TS. 1957 Dangers of cigarette-smoking. *BMJ* **2**, 297–298.
79. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. 1959 Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* **22**, 173–203.
80. Rosenbaum PR, Rubin DB. 1983 Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Stat. Soc. B* **45**, 212–218.
81. Ichino A, Mealli F, Nannicini T. 2008 From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity? *J. Appl. Econ.* **23**, 305–327. (doi:10.1002/jae.998)
82. Imbens GW. 2003 Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* **93**, 126–132. (doi:10.1257/000282803321946921)

83. Dorie V, Harada M, Carnegie NB, Hill J. 2016 A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Stat. Med.* **35**, 3453–3470. (doi:10.1002/sim.6973)
84. Ding P, VanderWeele TJ. 2016 Sensitivity analysis without assumptions. *Epidemiology* **27**, 368–377. (doi:10.1097/EDE.0000000000000457)
85. VanderWeele TJ, Ding P. 2017 Sensitivity analysis in observational research: introducing the E-value. *Ann. Intern. Med.* **167**, 268–274. (doi:10.7326/M16-2607)
86. Robins JM. 1999 Association, causation, and marginal structural models. *Synthese* **121**, 151–179. (doi:10.1023/A:1005285815569)
87. Rosenbaum PR. 2002 *Observational studies*. New York: Springer.
88. Rosenbaum PR, Silber JH. 2009 Amplification of sensitivity analysis in matched observational studies. *J. Am. Stat. Assoc.* **104**, 1398–1405. (doi:10.1198/jasa.2009.tm08470)
89. Angrist JD, Pischke J-S. 2009 *Mostly harmless econometrics: an empiricist's companion*. Princeton University Press.
90. Angrist JD, Imbens GW, Rubin DB. 1996 Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* **91**, 444–455. (doi:10.1080/01621459.1996.10476902)
91. Imbens GW, Rubin DB. 1997 Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Stat.* **25**, 305–327. (doi:10.1214/aos/1034276631)
92. Hirano K, Imbens GW, Rubin DB, Zhou XH. 2000 Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 69–88. (doi:10.1093/biostatistics/1.1.69)
93. Frangakis CE, Rubin DB. 2002 Principal stratification in causal inference. *Biometrics* **58**, 21–29. (doi:10.1111/j.0006-341X.2002.00021.x)
94. Frangakis CE, Rubin DB, Zhou XH. 2002 Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics* **3**, 147–164. (doi:10.1093/biostatistics/3.2.147)
95. Mealli F, Pacini B. 2013 Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *J. Am. Stat. Assoc.* **108**, 1120–1131. (doi:10.1080/01621459.2013.802238)
96. Zeng S, Li F, Ding P. 2020 Is being an only child harmful to psychological health?: evidence from an instrumental variable analysis of China's one-child policy. *J. R. Stat. Soc. A* **15**, 1615–1635. (doi:10.1111/rssa.12595)
97. Heckman JJ, Vytlacil EJ. 1999 Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proc. Natl Acad. Sci. USA* **96**, 4730–4734. (doi:10.1073/pnas.96.8.4730)
98. Zhang JL, Rubin DB, Mealli F. 2008 Evaluating the effects of job training programs on wages through principal stratification. In *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, pp. 117–145. New York, NY: John Wiley & Sons.
99. Gilbert PB, Hudgens MG. 2008 Evaluating candidate principal surrogate endpoints. *Biometrics* **64**, 1146–1154. (doi:10.1111/j.1541-0420.2008.01014.x)
100. Jiang Z, Ding P, Geng Z. 2016 Principal causal effect identification and surrogate end point evaluation by multiple trials. *J. R. Stat. Soc. B* **78**, 829–848. (doi:10.1111/rssb.12135)
101. Li F, Mattei A, 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. (doi:10.1214/15-AOAS881)
102. Ricciardi F, Mattei A, Mealli F. 2020 Bayesian inference for sequential treatments under latent sequential ignorability. *J. Am. Stat. Assoc.* **115**, 1498–1517. (doi:10.1080/01621459.2019.1623039)
103. Robins JM. 1986 A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Math. Modell.* **7**, 1393–1512. (doi:10.1016/0270-0255(86)90088-6)
104. Robins JM, Hernán MA, Brumback B. 2000 Marginal structural models and causal inference. *Epidemiology* **11**, 550–560. (doi:10.1097/00001648-200009000-00011)
105. Zajonc T. 2012 Bayesian inference for dynamic treatment regimes: mobility, equity, and efficiency in student tracking. *J. Am. Stat. Assoc.* **107**, 80–92. (doi:10.1080/01621459.2011.643747)
106. Gustafson P. 2015 Discussion of 'Bayesian estimation of marginal structural models' by Saarela *et al*. *Biometrics* **71**, 291–293. (doi:10.1111/biom.12271)

107. Robins JM, Wasserman L. 1997 Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proc. of the Thirteenth Conf. on Uncertainty in Artificial Intelligence*, pp. 409–420. Burlington, MA: Morgan Kaufmann Publishers Inc.
108. Saarela O, Stephens DA, Moodie EEM, Klein MB. 2015 On Bayesian estimation of marginal structural models. *Biometrics* **71**, 279–301. (doi:10.1111/biom.12269)
109. Chakraborty B, Murphy SA. 2014 Dynamic treatment regimes. *Ann. Rev. Stat. Appl.* **1**, 447–464. (doi:10.1146/annurev-statistics-022513-115553)
110. Scott EL. 2010 A modern Bayesian look at the multi-armed bandit. *Appl. Stoch. Models Bus. Ind.* **26**, 639–658. (doi:10.1002/asmb.874)
111. Ghavamzadeh M, Mannor S, Pineau J, Tamar A. 2015 Bayesian reinforcement learning: a survey. *Found. Trends Mach. Learn.* **8**, 359–483. (doi:10.1561/2200000049)
112. Schnell PM, Papadogeorgou G. 2020 Mitigating unobserved spatial confounding when estimating the effect of supermarket access on cardiovascular disease deaths. *Ann. Appl. Stat.* **14**, 2069–2095. (doi:10.1214/20-AOAS1377)
113. Daniels MJ, Roy JA, Kim C, Hogan JW, Perri MG. 2012 Bayesian inference for the causal effect of mediation. *Biometrics* **68**, 1028–1036. (doi:10.1111/j.1541-0420.2012.01781.x)
114. Forastiere L, Mealli F, Wu A, Airoldi E. 2022 Estimating causal effects under interference using Bayesian generalized propensity scores. *J. Mach. Learn. Res.* **23**, 1–61.
115. Zeng S, Rosenbaum S, Alberts SC, Archie EA, Li F. 2021 Causal mediation analysis for sparse and irregular longitudinal data. *Ann. Appl. Stat.* **15**, 747–767. (doi:10.1214/20-AOAS1427)
116. Stan Development Team. 2022 RStan: the R interface to Stan. R package version 2.21.5.