Statistical Science
2023, Vol. 38, No. 3, 369–385
https://doi.org/10.1214/22-STS879

The Role of Exchangeability in Causal Inference

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

Abstract. Though the notion of exchangeability has been discussed in the causal inference literature under various guises, it has rarely taken its original meaning as a symmetry property of probability distributions. As this property is a standard component of Bayesian inference, we argue that in Bayesian causal inference it is natural to link the causal model, including the notion of confounding and definition of causal contrasts of interest, to the concept of exchangeability. Here, we propose a probabilistic between-group exchangeability property as an identifying condition for causal effects, relate it to alternative conditions for unconfounded inferences (commonly stated using potential outcomes) and define causal contrasts in the presence of exchangeability in terms of posterior predictive expectations for further exchangeable units. While our main focus is on a point treatment setting, we also investigate how this reasoning carries over to longitudinal settings.

Key words and phrases: Bayesian inference, causal inference, confounding, exchangeability, posterior predictive inference.

1. INTRODUCTION

The concept of exchangeability has profound philosophical meaning in Bayesian statistics. Recall that an infinite sequence of observable random variables $(Y_i)_{i=1}^{\infty}$ is *exchangeable* if, for all finite n,

$$\begin{aligned}\Pr(Y_1 = y_1, \dots, Y_n = y_n) \\ = \Pr(Y_1 = y_{\rho(1)}, \dots, Y_n = y_{\rho(n)}),\end{aligned} \tag{1.1}$$

or $(Y_1, \dots, Y_n) \overset{\mathrm{d}}{=} (Y_{\rho(1)}, \dots, Y_{\rho(n)})$, for any permutation $\rho(\cdot)$ of the indices. This simple probabilistic definition plays a central, even totemic, role in Bayesian inference; it leads to the definition of "parameters" as functions of infinite sequences of observable quantities through de Finetti's representation theorem (de Finetti, 1929; a review of the original work is provided, e.g., by von Plato, 1989). This further facilitates probability statements on future, unobserved quantities based on information contained in observed data, and justifies the use of the posterior distribution as the basis for statistical inference (e.g., Bernardo and Smith, 1994, p. 173). In recent years, the term "exchangeability", or "conditional exchangeability", has been increasingly used in the field of causal inference. However, it has acquired a specific meaning synonymous with part of the *ignorability* assumption as stated by Rosenbaum and Rubin (1983), that is, a certain conditional independence relationship between exposure (or treatment), potential outcomes and possible confounding variables. In this paper, we study the links between the two usages of the term and point out their common underlying probabilistic arguments. Furthermore, we propose a fully Bayesian formulation of causal inference that is based on exchangeable representations and includes Bayesian definitions of causal estimands. Our central thesis is that de Finetti's formulation of exchangeability is entirely sufficient to give a coherent basis for causal inference, without the need to introduce special constructs (such as potential outcomes), mathematical machinery (such as the *do*-operator) or additional conditional independence assumptions.

1.1 Review of the de Finetti Representation for Exchangeable Sequences and the Problem Setup

The de Finetti representation theorem for exchangeable sequences is a key mathematical result, which underpins all Bayesian inference methodology. The original version

Olli Saarela is Associate Professor, Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, Ontario M5T 3M7, Canada (e-mail: olli.saarela@utoronto.ca). David A. Stephens is Professor, Department of Mathematics and Statistics, McGill University, Burnside Hall, 805 Sherbrooke Street West, Montreal, Quebec H3A 0B9, Canada (e-mail: d.stephens@math.mcgill.ca). Erica E. M. Moodie is Professor, Department of Epidemiology and Biostatistics, McGill University, 2001 McGill College Ave, Montreal, Quebec H3A 1G1, Canada (e-mail: erica.moodie@mcgill.ca).

for binary sequences was generalized to any real-valued random quantities by Hewitt and Savage (1955), and the generalized version has been restated, for example, as Proposition 4.3 of Bernardo and Smith (1994). This states that if $(Y_i)_{i=1}^{\infty}$ is an infinite exchangeable sequence of random variables with probability law Pr, there exists a random probability measure, P, such that conditionally on P, the Y_n are independent and identically distributed (i.i.d.) with common distribution P. Moreover, with probability one, such a P is the weak limit of the sequence of empirical distributions $\hat{P}_n(B) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i \in B\}}$, where $B \subseteq \mathbb{R}$. In Bayesian learning under exchangeability, the random probability measure P can be interpreted as an infinite-dimensional "parameter," with probability law, say Q, interpreted as the "prior" belief distribution. Hierarchically, this means that $Y_i \mid P \sim_{\text{i.i.d.}} P$ and $P \sim Q$. In the notation that follows, we distinguish between the "marginal" measure, Pr, and the random "parameter-conditional" measure, P, as the latter's existence is implied by the exchangeability property on the marginal distribution.

To characterize the conditional probability structures that appear in causal settings, we need the notion of *partial exchangeability* originally introduced by de Finetti (1938) and reviewed, for example, by Diaconis (1988). Partial exchangeability characterizes the comparability of units within subpopulations that are formed, for example, by a (categorical) covariate. In our most basic setting, we have $W_i = (Y_i, Z_i, X_i)$ where Y_i is an observable outcome, Z_i is an observable treatment/exposure and X_i represents (typically a vector of) possible confounding variables. For simplicity, we consider the case where all variables take a finite number of possible values, possibly after discretizing continuous variables, so that $Y_i \in \mathcal{Y} \equiv \{0, 1, \dots, \ell\}$, $Z_i \in \mathcal{Z} \equiv \{0, 1, \dots, m\}$ and $X_i \in \mathcal{X} \equiv \{0, 1, \dots, p\}$. However, we note that it is straightforward to generalize everything that follows to continuous outcomes Y_i (see, e.g., Definition 4.14 of Bernardo and Smith, 1994, for a generalization based on unrestricted exchangeability for sequences with predictive sufficient statistics).

For the joint distribution, for any $n \geq 1$ and combination of values $z_i \in \mathcal{Z}$, $x_i \in \mathcal{X}$ with a positive probability, we have the factorization

$$\begin{aligned}
&\Pr\left(\bigcap_{i=1}^{n}(Y_i = y_i, Z_i = z_i, X_i = x_i)\right)\\
&\quad = \Pr\left(\bigcap_{i=1}^{n}(Y_i = y_i) | \bigcap_{i=1}^{n}(Z_i = z_i, X_i = x_i)\right)\\
&\qquad \times \Pr\left(\bigcap_{i=1}^{n}(Z_i = z_i) | \bigcap_{i=1}^{n}(X_i = x_i)\right)\\
&\qquad \times \Pr\left(\bigcap_{i=1}^{n}(X_i = x_i)\right).
\end{aligned}$$

Assuming exchangeability of the random vectors W_i over the individual indices i, identity (1.1) becomes

$$\begin{aligned}
&\Pr\left(\bigcap_{i=1}^{n}(Y_i = y_i, Z_i = z_i, X_i = x_i)\right)\\
&\quad = \Pr\left(\bigcap_{i=1}^{n}(Y_i = y_{\rho(i)}, Z_i = z_{\rho(i)}, X_i = x_{\rho(i)})\right),
\end{aligned}$$

where ρ permutes the individual indices. By considering permutations ρ that preserve the values of Z and X (so that $z_{\rho(i)} = z_i$ and $x_{\rho(i)} = x_i$), the exchangeability over i also implies that

$$\begin{aligned}
&\Pr\left(\bigcap_{i=1}^{n}(Y_i = y_i) \,\middle|\, \bigcap_{i=1}^{n}(Z_i = z_i, X_i = x_i)\right)\\
&\quad = \Pr\left(\bigcap_{z,x}\bigcap_{i \in I_{zx}^{n}}(Y_i = y_{\rho_{zx}(i)}) \,\middle|\, \bigcap_{i=1}^{n}(Z_i = z_i, X_i = x_i)\right),
\end{aligned} \tag{1.2}$$

where $(z, x) \in \mathcal{Z} \times \mathcal{X}$ and ρ_{zx} permutes the indices within the index set $I_{zx}^{n} = \{1, \dots, n\} \cap \{i : Z_i = z, X_i = x\}$. Identity (1.2) corresponds to de Finetti's definition of partial exchangeability and, for example, in the case of $\ell = 1$, implies the joint representation

$$\begin{aligned}
&\Pr\left(\bigcap_{i=1}^{n}(Y_i = y_i) \,\middle|\, \bigcap_{i=1}^{n}(Z_i = z_i, X_i = x_i)\right)\\
&\quad = \int_{\mathcal{P}} \prod_{z,x} \prod_{i \in I_{zx}^{n}} P(Y_i = y_i \mid Z_i = z,\\
&\qquad X_i = x; \phi_{zx})\, \mathrm{d}Q(\phi),
\end{aligned} \tag{1.3}$$

where $\phi = (\phi_{00}, \dots, \phi_{mp})$, $P(Y_i = y_i \mid Z_i = z, X_i = x; \phi_{zx}) = \phi_{zx}^{y_i}(1 - \phi_{zx})^{1-y_i}$,

$$Q(\phi) = \lim_{n \longrightarrow \infty} \Pr\left\{\bigcap_{z,x}\left(\frac{\sum_{i=1}^{n} \mathbf{1}_{\{Y_i=1, Z_i=z, X_i=x\}}}{\sum_{i=1}^{n} \mathbf{1}_{\{Z_i=z, X_i=x\}}} \leq \phi_{zx}\right)\right\},$$

and

$$\phi_{zx} = \lim_{n \longrightarrow \infty} \frac{\sum_{i=1}^{n} \mathbf{1}_{\{Y_i=1, Z_i=z, X_i=x\}}}{\sum_{i=1}^{n} \mathbf{1}_{\{Z_i=z, X_i=x\}}}.$$

The interpretation of (1.3) is that within each treatment/covariate stratum the outcomes are conditionally independent and distributed as $Y_i \mid (Z_i = z, X_i = x; \phi_{zx}) \sim$ Bernoulli(ϕ_{zx}), and Q, which is a multivariate cumulative distribution function, is the prior belief distribution on the long-run, stratum-specific relative frequencies. Another interpretation is that the stratum-specific event counts are sufficient statistics with binomial distributions. The model specification would be completed by the specification of Q; a full discussion of the prior specification is beyond our scope here, but we note two special cases. Assuming $\phi_{00} = \cdots = \phi_{mp}$ would imply the exchangeability of the entire sequence (no difference between the groups), whereas assuming the group-specific parameters ϕ_{zx}s themselves to be exchangeable would imply a hierarchical form for the representation (see, e.g., Section 4.6.5 of Bernardo and Smith, 1994). We note that the latter property is different from the between-group exchangeability that we introduce in Section 3 for causal considerations.

While representations such as (1.3) enable statistical inferences on the unobservable characteristics of the infinite sequences based on observable finite sequences, further assumptions are needed for causal interpretations. Consider, for example, the case of $m = 1$, with $Z_1 = 1$ and $Z_1 = 0$ representing the intervention and control groups, respectively. Here, the covariate stratum-specific risk differences $\phi_{1x} - \phi_{0x}$ or risk ratios ϕ_{1x}/ϕ_{0x}, or their marginal counterparts based on standardized risks $\sum_x \phi_{zx} P(X_i = x)$, would not be causal contrasts without further assumptions on the treatment assignment mechanism. We will argue that ruling out unmeasured confounding requires a specific kind of between-group exchangeability in addition to the within-group property stated in (1.2).

1.2 Literature Review: Exchangeability and Causal Inference

A connection between the original probabilistic concept of exchangeability and causal inference was first suggested by Lindley and Novick (1981, p. 51); however, the authors did not pursue this further. This connection was pointed out later by Greenland and Robins (1986) in the context of nonidentifiability of causal parameters due to confounding. However, in the causal inference literature (e.g., Greenland, Robins and Pearl, 1999, Hernán and Robins, 2006, Greenland and Robins, 2009), exchangeability has been interpreted in terms of potential outcomes (instead of observable quantities), and the connection of this concept to its Bayesian interpretation appears to have been lost. In this paper, we highlight the similarities between causal reasoning based on unit-level exchangeability and the now more common formulation based on potential outcomes.

We aim to provide a sequel to the classic account of Lindley and Novick (1981) that takes into account the numerous developments that have taken place in causal inference theory and methodology since. The utility of the concept of exchangeability and the account of Lindley and Novick (1981) have been disputed by Pearl (2009, pp. 177–180) (see also Lindley, 2002), who argued that probability theory alone is not adequate for providing a comprehensive framework for causal reasoning (which, in fact, Lindley and Novick never attempted). Rather than enter this debate, we concentrate on clarifying the connection between the probabilistic notion of exchangeability and causal inference, using exchangeability as the basis of the *causal model.* A causal model is necessary to define the causal contrast of interest, as well as to define the notion of confounding and to state the identifying conditions required for unconfounded inferences.

We follow the key insight of Lindley and Novick (1981, p. 45) that "inference is a process whereby one passes from data on a set of units to statements about a further unit." Because we can only ever observe outcomes for any individual unit under a single exposure pattern, it seems reasonable to base statistical inferences about causal effects on an explicit assumption of "similarity" (or more precisely, indistinguishability) of the individual instances. To assume an exchangeable structure is always appropriate after sufficient relevant information has been included (Gelman et al., 2004, p. 6); however, what constitutes sufficient relevant information in causal inference settings often has to be decided based on prior information alone, as noted by Greenland and Robins (2009). That is, causal inferences from observational settings necessarily rely on prior information regarding the causal mechanisms involved; the role of prior information can be made explicit in Bayesian inference.

Several other authors have attempted to make connections between classical statistical models and causal models. In particular, Dawid (2000), Arjas and Parner (2004) and Chib (2007) have suggested that the potential outcomes notation is redundant in formulating causal models, and similar arguments have been made both in Bayesian and frequentist settings. Baker (2013) gave a probabilistic interpretation to confounding and collider biases. Many of the formulations put forth as alternatives to potential outcomes are based on introducing a hypothetical "randomized" or "experimental" probability measure that is used to formulate the causal quantity of interest (Dawid and Didelez, 2010, Røysland, 2011, Arjas, 2012, Saarela et al., 2015, Commenges, 2019). Inference then becomes a matter of linking the experimental measure to the observational one thought to have generated the data, which involves assumptions about the absence of unmeasured confounding. Other formulations are based on structural definitions, where a deterministic relationship is assumed between observed and latent variables (Commenges and Gégout-Petit, 2015, Ferreira, 2019).

The "no confounding" assumption required for identification of the causal effect under these formulations is usually expressed in terms of latent variables, or equivalence of certain components of the experimental and observational joint distributions, termed by Dawid and Didelez (2010) as the *stability* assumption. Bühlmann (2020) termed a similar property "invariance" and formulated causal inference in terms of a risk minimization problem. Ferreira (2015) framed an exchangeability property concerning the treatment assignment mechanism as a "no confounding" type assumption, but they did not connect it to Bayesian inference. We are not aware of exchangeability (in its original meaning as a symmetry property of probability distributions) otherwise used as a causal assumption; Dawid, Musio and Fienberg (2016) used it as an inferential assumption needed in addition to a "no confounding" type assumption. Dawid (2021) made a distinction between post-treatment and pretreatment exchangeability, where the latter is closely related to the notion of partial exchangeability of outcomes within treatment and control groups separately, while the former involves a judgment of similarity of the groups being compared before they received treatment. A further ignorability condition concerning the treatment assignment mechanism is needed for causal inferences based on the observed responses in the treatment and control groups.

Like Dawid (2021), we consider partial exchangeability, as defined above, as a starting point, suggesting parametric inferences based on the representation theorem. However, while this within-group exchangeability is sufficient for predicting the outcome for a further exchangeable unit, causal inferences require a judgment on exchangeability *between* groups, that is, between treated and untreated units, reflecting the absence of confounding due to the group characteristics. In this work, our primary objective is to formulate the required condition as a probabilistic symmetry property. Furthermore, we show that this property indeed is an identifiability condition for causal effects as it implies ignorability of the treatment assignment mechanism. Under this condition, the parameters suggested by the representation theorem have a causal interpretation, which provides a link to Bayesian causal inferences. We further extend this reasoning to longitudinal settings, where in addition to biases due to confounding, we can encounter biases related to conditioning on intermediate variables. Similar to Ferreira (2019), we adopt a structural model notation as this allows us to draw connections between the different causal models but with a focus on Bayesian causal inference.

1.3 Manuscript Outline

The paper proceeds as follows. In Section 2, we introduce the necessary notation and concepts. In Section 3, we propose a definition of conditional exchangeability to be used as an identifying condition for estimating causal effects. We show that this condition implies ignorability of the treatment assignment mechanism and relate it to alternative conditions based on causal diagrams and potential outcomes. In Section 4, we give a Bayesian definition of a marginal causal contrast and consider inference under observational settings. In Section 5, we consider extending the proposed framework to longitudinal settings. We conclude with a discussion in Section 6.

2. NOTATION AND FOUNDATIONS

2.1 Structural Assumption

It is convenient for our derivations to assume that the outcome random variable, Y_i, is determined by the structural rule $Y_i = f(Z_i, X_i, U_i)$, where Z_i represents treatment assignment, X_i observed potential confounders and U_i unobserved factors that may be determinants of Y_i and may or may not also be confounders. This structural assumption is quite general as the model can be readily modified to include further stochastic elements such as additive "residual" errors. Note that in the structural definition, we may consider specific *interventions* on treatment and write $f(z, X_i, U_i)$, as if random variable Z_i has a degenerate distribution at z, and so that the intervention is independent of (X_i, U_i). Note also that the structural definition is essentially identical to the potential outcome construction; in the conventional notation, the potential outcome is given by $Y_i(z) \equiv f(z, X_i, U_i)$. In what follows, we always assume "general" infinite exchangeability of the sequence $((Z_i, X_i, U_i))_{i=1}^{\infty}$ (and consequently $(Y_i)_{i=1}^{\infty}$ as it is determined by the former) over the individual indices i, which also implies exchangeability of the sequence $(W_i)_{i=1}^{\infty}$ of the observable random vectors $W_i = (Y_i, Z_i, X_i)$. For finite sequences of these, in places we use vector notation $(W_i)_{i=1}^{n} = (W_1, \dots, W_n)$.

2.2 Experimental and Observational Designs

The objective of causal inference is to quantify the effect of assigning a treatment level, z, (relative to an alternative level z') on the outcome, independent of any other determinants of the outcome. Such an allocation mechanism is commonly termed *experimental*. We label the corresponding probability distributions of observations under such a setting by $\mathcal{E}$. If the independence is not known to be present, the mechanism is termed *observational*, or *non-experimental*. The corresponding distributions are labeled by $\mathcal{O}$. The independence requirement may be expressed as the factorization

$$\begin{aligned} &\Pr\Big(\bigcap_{i=1}^{n}(Z_i = z_i, X_i = x_i, U_i \in \mathrm{d}u_i; \mathcal{E}\Big) \\ &= \Pr\Bigg(\bigcap_{i=1}^{n}(Z_i = z_i); \mathcal{E}\Bigg) \\ &\quad \times \Pr\Bigg(\bigcap_{i=1}^{n}(X_i = x_i, U_i \in \mathrm{d}u_i); \mathcal{E}\Bigg), \end{aligned} \tag{2.1}$$

for any $n \geq 1$, where each of the factors on the right-hand side has a representation of the form of (1.3). From this, it also follows that $Z_j \perp\!\!\!\perp (X_k, U_k)$ for all j, k, where we use $\perp\!\!\!\perp$ to denote statistical independence. This expression could be generalized to allow the treatment assignment to depend on the observed characteristics X_i, but in what follows we proceed with (2.1).

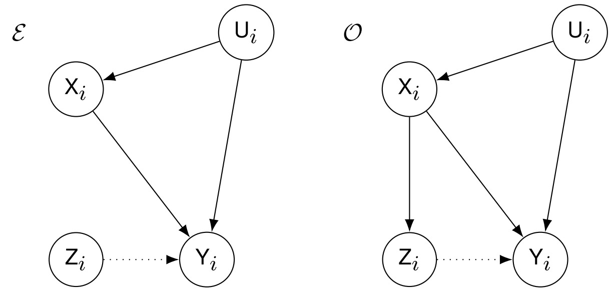

FIG. 1. *Left-hand panel*: *DAG depicting a randomized setting labeled $\mathcal{E}$. The dotted arrow $Z_i \longrightarrow Y_i$ is absent if there is no treatment effect. Right-hand panel*: *DAG depicting an observational setting labeled $\mathcal{O}$. The arrow $X_i \longrightarrow Z_i$ is the distinguishing feature of $\mathcal{O}$ compared to $\mathcal{E}$; conditioning on Z_i would open a confounding "backdoor" path from Z_i to Y_i.*

2.3 Directed Acyclic Graphs

In subsequent sections, our explanations are assisted by the use of directed acyclic graphs (DAGs) to illustrate the underlying relationships between the variables. In the Bayesian framework, we can regard a posited DAG as encapsulating structural prior knowledge related to the observable quantities, and they may be considered either conditional on or marginalized over parameters in models. In this paper, we use the terms "knowledge," "information" and "opinion" interchangeably to describe the a priori-held subjective beliefs—both qualitative and quantitative—of the experimenter. As a notational device, we will use structural definitions to illustrate the link between the information encoded in a DAG and the corresponding probability statements.

The DAG in the left-hand panel of Figure 1 illustrates the relationship between variables as described in Section 2.1 and where Z is assigned experimentally. Figure 1 relates to a single individual i; by convention, in the frequentist setting, the nodes on a single DAG are interpreted to indicate probabilistic relationships for random variables relating to an archetypal individual present in a random sample, with the graph replicated identically across the independent draws $i = 1, \dots, n$. As indicated by equation (1.3), however, under an assumption of exchangeability, the collections of variables $W_i, i = 1, \dots, n$ are not marginally independent, but instead are conditionally independent given parameter P. Under the assumption of exchangeability of the W_i, the most general DAG would have an additional node containing P from which arrows into the complete collection of variables would emanate (Figure A1 in Supplementary Appendix A; Saarela, Stephens and Moodie, 2023).

3. EXCHANGEABILITY AND IGNORABILITY

3.1 Exchangeability Under Randomization

Under the randomized setting $\mathcal{E}$, factorization (2.1) and the general exchangeability of (X_i, U_i) imply an additional exchangeability property that we can give a causal interpretation. A similar property can then be considered as an identifying assumption for causal effects in an observational setting $\mathcal{O}$, where this property is not implied by design. Now, taking $A \equiv \{Z_j = z, Z_k = z'\}$ to be the observed treatment assignment,

$$\begin{aligned}
&\Pr(Y_j = y, Y_k = y' \mid A; \mathcal{E}) \\
&\quad = \Pr(f(z, X_j, U_j) = y, f(z', X_k, U_k) = y' \mid A; \mathcal{E}) \\
&\quad = \Pr(f(z, X_j, U_j) = y, f(z', X_k, U_k) = y'; \mathcal{E}) \qquad (3.1)\\
&\quad = \Pr(f(z, X_k, U_k) = y, f(z', X_j, U_j) = y'; \mathcal{E}) \\
&\quad = \Pr(f(z, X_k, U_k) = y, f(z', X_j, U_j) = y' \mid A; \mathcal{E})
\end{aligned}$$

for all (y, y') and (z, z'). Here, the first equality followed from the functional definition, third from exchangeability and second and fourth from independence of the assignment mechanism. In particular, (3.1) states that under the experimental setting, the joint distribution of the two outcomes is the same under a hypothetical switch of the interventions. Thus, taking $z = 1$ and $z' = 0$ and $A \equiv \{Z_j = 1, Z_k = 0\}$, the property

$$\begin{aligned}
&\Pr(f(1, X_j, U_j) = y, f(0, X_k, U_k) = y' \mid A; \mathcal{E}) \\
&\quad = \Pr(f(1, X_k, U_k) = y, f(0, X_j, U_j) = y' \mid A; \mathcal{E})
\end{aligned}$$

suggests a causal interpretation; the joint distribution of the outcomes does not depend on which individual was actually assigned treatment $z = 1$. In other words, the known treatment assignment A is not informative of the other determinants of the outcomes. This property does not follow from the previously assumed exchangeability over i,

$$\begin{aligned}
&\Pr(Y_j = y, Y_k = y' \mid Z_j = 1, Z_k = 0; \mathcal{E}) \\
&\quad = \Pr(Y_k = y, Y_j = y' \mid Z_j = 0, Z_k = 1; \mathcal{E}),
\end{aligned}$$

that is, even under the experimental setting, the statement

$$\begin{aligned}
&\Pr(Y_j = y, Y_k = y' \mid Z_j = 1, Z_k = 0; \mathcal{E}) \\
&\quad = \Pr(Y_k = y, Y_j = y' \mid Z_j = 1, Z_k = 0; \mathcal{E})
\end{aligned}$$

would only be true if there were no treatment effect. While we could consider such "under the null" causal exchangeability statements, the structural model allows us to make explicit the hypothetical switching of the treatments without this restriction.

Statement (3.1) can be extended to any finite sequence of observations, conditional on a sequence of treatment assignments, as

$$\begin{aligned}
&\Pr\left(\bigcap_{i=1}^{n} (f(z_i, X_i, U_i) = y_i) \,\middle|\, \bigcap_{i=1}^{n} (Z_i = z_i); \mathcal{E}\right) \\
&\quad = \Pr\left(\bigcap_{i=1}^{n} (f(z_i, X_{\rho(i)}, U_{\rho(i)}) = y_i) \,\middle|\, \bigcap_{i=1}^{n} (Z_i = z_i); \mathcal{E}\right) \qquad (3.2)
\end{aligned}$$

for any permutation $\rho(\cdot)$ of the indices. In the remainder of this section, we show that under property (3.2), the parameters implied by representation (1.3), such as contrasts of treatment group specific outcome frequencies/risks, have a causal interpretation. We note first that by (3.2), considering permutations only within the treatment groups, the sequences of "treated" random variables $f(1, X_i, U_i)$ and "untreated" random variables $f(0, X_i, U_i)$ are partially exchangeable. Thus, the within-group exchangeability of outcome sequences (1.2) is a special case of (3.2), the interpretation being that the stronger condition extends partial exchangeability to certain kinds of between-group comparisons. Essentially, (3.2) states that the remaining determinants, observed and unobserved, of the outcomes are exchangeable between the treatment groups. We note that (3.2) would follow from assuming X_i and U_i to be similarly exchangeable, but this would be an unnecessarily strong assumption, as in (3.2) this is only required for the aspects of X_i and U_i that are determinants of the outcome.

We term property (3.1) and its extension (3.2) as *conditional exchangeability* to distinguish them from the previously assumed partial exchangeability. While under the experimental setting these were implied by the latter and the known properties of the treatment assignment mechanism, under observational settings considered in Section 3.2, a similar property will have to be assumed a priori. When the assignment mechanism is unknown, this is a strong assumption, but one that is needed for the identifiability of causal effects based on observational studies. It then becomes important that the conditional exchangeability statements imply properties of the treatment assignment mechanism. To see this, from (3.1) it follows that

$$\begin{aligned}
&\sum_{y'} \Pr\bigl(f(1, X_j, U_j) = y, f(0, X_k, U_k) = y' \mid A; \mathcal{E}\bigr) \\
&\quad = \sum_{y'} \Pr\bigl(f(1, X_k, U_k) = y, f(0, X_j, U_j) = y' \mid A; \mathcal{E}\bigr) \\
&\quad \Rightarrow \quad \Pr\bigl(f(1, X_j, U_j) = y \mid A; \mathcal{E}\bigr) \\
&\qquad\qquad = \Pr\bigl(f(1, X_k, U_k) = y \mid A; \mathcal{E}\bigr),
\end{aligned}$$

that is, $f(1, X_j, U_j) \mid (Z_j = 1, Z_k = 0) \overset{\mathrm{d}}{=} f(1, X_k, U_k) \mid (Z_j = 1, Z_k = 0)$ under $\mathcal{E}$. If we further assume that the treatment assignment of individual k is not informative of the outcome of individual j and vice versa (corresponding to the common assumption of "no interference between units," cf. Rubin, 1978; Lindley and Novick, 1981, p. 58) we have that

$$f(1, X_j, U_j) \mid (Z_j = 1) \overset{\mathrm{d}}{=} f(1, X_k, U_k) \mid (Z_k = 0).$$

Further, by general exchangeability we have that

$$f(1, X_j, U_j) \mid (Z_j = 1) \overset{\mathrm{d}}{=} f(1, X_k, U_k) \mid (Z_k = 1),$$

and combining this with the previous, that $f(1, X_i, U_i) \perp\!\!\!\perp Z_i$ under $\mathcal{E}$. By a symmetrical argument, we can show that $f(0, X_i, U_i) \perp\!\!\!\perp Z_i$, and finally that $f(z, X_i, U_i) \perp\!\!\!\perp Z_i, z \in \{0, 1\}$ in the Pr distribution. This independence property was implied by (3.1) and the "no interference between units" assumption.

While the previous applies marginally, symmetry property (3.2) holds true also conditional on the parameters implied by the de Finetti representation, following the arguments in the Appendix. We also note that in the P distribution, the "no interference between units" property is implied by the general exchangeability due to the resulting i.i.d. structure. Thus, we also have $f(z, X_i, U_i) \perp\!\!\!\perp Z_i$ in the P distribution, which is equivalent to the ignorability condition $Y_i(z) \perp\!\!\!\perp Z_i$ commonly stated in terms of potential outcomes. We return to this connection in Section 3.6 but note that under the randomized setting, we have demonstrated that exchangeability and ignorability both express a similar "no confounding" property. This property allows for unconfounded comparisons of the treatment arms in terms of long-run outcome frequencies. Expressing this as a probabilistic symmetry statement allows us to make use of Bayesian concepts in outlining a causal modeling framework. A perceived strength of the potential outcomes framework is being able to express causal contrasts of interest directly in terms of the average potential outcomes, such as $E[Y(1)] - E[Y(0)]$ and $E[Y(1)]/E[Y(0)]$ for risk difference and ratio, respectively, without referring to parameters in statistical models. Similar constructs are also possible in the present framework, which we will address in Section 4.

3.2 Exchangeability in the Observational Setting

While it was helpful to demonstrate the ideal properties of the experimental setting, we are actually interested in inferences under observational settings, where we do not choose the treatment assignment mechanism and the exchangeability of the subpopulations being compared does not follow from the study design. We consider a hypothetical study of the effect of initiation of antiretroviral therapy on CD4 cell counts, based on a cohort of n HIV patients. Specifically, for $i = 1, \dots, n$, let random variable X_i represent a baseline CD4 cell count measurement for HIV-positive individual i, Z_i represent the decision to initiate antiretroviral therapy at the baseline time point, and Y_i the CD4 cell count measurement after a fixed time has passed since baseline. Further, let U_i be a latent variable representing the underlying immune status of individual i, some facet of which could possibly be captured by X_i. We know that individuals with lower CD4 cell counts, X_i, are more likely to initiate treatment; the correlation between U_i and X_i further implies that those with weakened underlying immune function are more likely to initiate treatment. Thus, the factorization in (2.1) likely does not hold.

Such dependencies are illustrated in the right-hand DAG in Figure 1.

If we are interested in the causal effect of treatment initiation, it seems appropriate to formulate the *causal estimand* of interest in terms of a hypothetical randomized trial that otherwise resembles the observational setting but where (2.1) holds. The structural assumption $Y_i = f(Z_i, X_i, U_i)$ can be taken to apply under both settings, and the distribution of (X_i, U_i) can also be assumed to be the same as we don't actually observe any data under $\mathcal{E}$. The problem of causal inference then involves making probability statements about the estimand specified under $\mathcal{E}$ based on data collected under $\mathcal{O}$. Based on the context, a simple comparison in terms of the summary statistics of the outcomes among those observed to be treated and those observed not would be *confounded*; a numerical example demonstrating this is presented in Supplementary Appendix B. We formalize this concept in Section 4, where we introduce an explicit causal estimand and its estimator. Before that, we attempt to understand this noncomparability of the groups through a probabilistic exchangeability statement between individuals representative of those groups and propose this statement as an identifying condition for causal effects.

Consider a comparison of two individuals, j and k, with observed treatment assignments $Z_j = 1$ and $Z_k = 0$, respectively, but with the outcome yet to be observed. As in Section 3.1, we consider whether the outcomes of these two individuals are exchangeable (pairwise) under a hypothetical intervention to reverse their treatments; if so, a comparison of their outcomes under the actual treatment assignments would be informative of the causal effect of the treatment. We again take the outcome to be determined by a structural model, in which case the required exchangeability property conditional on $A \equiv \{Z_j = 1, Z_k = 0\}$ can be expressed as

$$\begin{aligned} &\Pr\bigl(Y_j = y, Y_k = y' \mid A; \mathcal{O}\bigr) \\ (3.3)\quad &= \Pr\bigl(f(1, X_j, U_j) = y, f(0, X_k, U_k) = y' \mid A; \mathcal{O}\bigr) \\ &= \Pr\bigl(f(0, X_j, U_j) = y', f(1, X_k, U_k) = y \mid A; \mathcal{O}\bigr) \end{aligned}$$

for all (y, y'), which mirrors the property obtained under $\mathcal{E}$. We emphasize that a statement such as (3.3) could usually only be made on a subjective basis, conditional on information concerning the study design and data generating mechanism; it represents a strong assumption requiring no unmeasured confounding. In addition, the causal question of interest, including the role of the variables in the data generating mechanism, must be stated a priori; without this knowledge, we would not know which exchangeability judgment is relevant for drawing causal inferences. Identity (3.3) could be extended to any finite sequence similar to (3.2). If this property holds under $\mathcal{O}$, and we additionally assume no interference between units, we would obtain $Z_i \perp\!\!\!\perp f(z, X_i, U_i)$ under $\mathcal{O}$.

Central to (3.3) for causal considerations is the extent to which group assignment can tell us about the other characteristics of the groups through the a priori knowledge of the relationships between the variables. If statement (3.3) was true, the treatment and reference groups, and individuals j and k, would be directly comparable, implying that a comparison of the two groups through a suitable summary statistic, for instance,

$$\frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i} - \frac{\sum_{i=1}^n (1 - z_i) y_i}{\sum_{i=1}^n (1 - z_i)}$$

would be free from confounding. However, exchangeability of the units of inference implies that the *labels of the units do not carry relevant information* (e.g., Bernardo and Smith, 1994, p. 168; Gelman et al., 2004, p. 6), which is now clearly not the case because of how the comparison was constructed: a priori we would expect individual j to have lower baseline CD4 count than k based on the treatment assignments.

3.3 Restoring Exchangeability Through Conditioning

The strong assumption in (3.3) can be weakened using conditioning. In the example above, if the baseline CD4 count sufficiently represents the indication to initiate treatment, we can stratify on this variable to achieve better comparability. Let now j and k index treated $(Z_j = 1)$ and untreated $(Z_k = 0)$ individuals matched on the condition $X_j = X_k = x$. Now, we could assume conditional on $A_x \equiv \{Z_j = 1, Z_k = 0, X_j = X_k = x\}$ that

$$\begin{aligned} &\Pr\bigl(Y_j = y, Y_k = y' \mid A_x; \mathcal{O}\bigr) \\ (3.4)\quad &= \Pr\bigl(f(1, x, U_j) = y, f(0, x, U_k) = y' \mid A_x; \mathcal{O}\bigr) \\ &= \Pr\bigl(f(0, x, U_j) = y', f(1, x, U_k) = y \mid A_x; \mathcal{O}\bigr) \end{aligned}$$

for all (y, y'). Similar to the discussion in Section 3.1, (3.4) implies that $f(z, x, U_j) \mid (Z_j = 1, Z_k = 0, X_j = X_k = x) \overset{\mathrm{d}}{=} f(z, x, U_k) \mid (Z_j = 1, Z_k = 0, X_j = X_k = x)$ under $\mathcal{O}$. And further, under the assumption of no interference between the units, $Z_i \perp\!\!\!\perp f(z, x, U_i) \mid X_i = x$ under $\mathcal{O}$. Condition (3.4) can be extended to any finite sequence matched on x, similar to (3.2). Because (3.4) applies also under the experimental setting $\mathcal{E}$ and we assume the distribution of the baseline characteristics (X_i, U_i) to be the same in both $\mathcal{O}$ and $\mathcal{E}$, we also have that

$$\begin{aligned} &\Pr(Y_i \mid Z_i = z, X_i = x; \mathcal{E}) \\ &\quad = \Pr\bigl(f(z, x, U_i) \mid Z_i = z, X_i = x; \mathcal{E}\bigr) \\ &\quad = \Pr\bigl(f(z, x, U_i) \mid X_i = x; \mathcal{E}\bigr) \\ &\quad = \Pr\bigl(f(z, x, U_i) \mid X_i = x; \mathcal{O}\bigr) \\ &\quad = \Pr\bigl(f(z, x, U_i) \mid Z_i = z, X_i = x; \mathcal{O}\bigr) \\ &\quad = \Pr(Y_i \mid Z_i = z, X_i = x; \mathcal{O}). \end{aligned}$$

Following the arguments in the Appendix, the same property would also apply in the i.i.d. distribution implied by the infinite exchangeability. The corresponding equivalence

$$P(Y_i \mid Z_i, X_i; \mathcal{E}) = P(Y_i \mid Z_i, X_i; \mathcal{O})$$

is the "no confounding" condition termed stability by Dawid and Didelez (2010). Here, exchangeability implies stability for the conditional outcome distribution, which is one of the required identifying conditions for inferences on marginal causal contrasts (Section 4.2).

3.4 Connection to Posterior Predictive Inferences

Comparison (3.4) involved two individuals with an opposite treatment assignment and is the relevant comparison for causal considerations. For predictive considerations, the general infinite exchangeability is sufficient, implying the partial exchangeability of the outcomes within subgroups with the same characteristics, that is,

$$\begin{aligned}&\Pr\bigl(Y_j = y, Y_k = y' \mid Z_j = Z_k = z, X_j = X_k = x; \mathcal{O}\bigr)\\ (3.5)\quad &= \Pr\bigl(Y_j = y', Y_k = y \mid Z_j = Z_k = z,\\ &\qquad X_j = X_k = x; \mathcal{O}\bigr).\end{aligned}$$

The pairwise exchangeability statement (3.5) extends from observed units $i = 1, \dots, n$ to further similarly matched units j and k, $j, k > n$, which motivates the use of posterior predictive inferences within the treatment groups. The further conditional exchangeability consideration (3.4) suggests that the predictions can be compared across the treatment groups as

$$\begin{aligned}&\sum_x \bigl(E[Y_j \mid Z_j = 1, X_j = x, D_{1x}; \mathcal{O}]\\ &\qquad - E[Y_k \mid Z_k = 0, X_k = x, D_{0x}; \mathcal{O}]\bigr)\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i = x\}}\\ (3.6)\quad &\approx \sum_x \biggl(\frac{\sum_{i=1}^n \mathbf{1}_{\{X_i = x\}} z_i y_i}{\sum_{i=1}^n \mathbf{1}_{\{X_i = x\}} z_i}\\ &\qquad - \frac{\sum_{i=1}^n \mathbf{1}_{\{X_i = x\}}(1 - z_i) y_i}{\sum_{i=1}^n \mathbf{1}_{\{X_i = x\}}(1 - z_i)}\biggr)\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i = x\}},\end{aligned}$$

where $D_{zx} \equiv \{W_i : i \in I_{zx}^n\}$ denotes the observed data on the matched groups and where the last form follows by approximating the within-stratum posterior predictive means by the sample means (cf. Lindley and Novick, 1981, p. 47). Thus, the above recovers the classical direct standardization formula for the marginal treatment effect (Keiding and Clayton, 2014, also known as the backdoor adjustment formula, Pearl, 2009).

For Bayesian inference, if the strata are too small for the use of the direct standardization formula, one would instead have to pool the observed data and connect them to the predictions through parametric probability models. We will formalize this in the following section but note here that the modeling approach requires the existence of parameter vectors Φ and Ψ given which $Y_j \perp\!\!\!\perp (W_i)_{i=1}^n \mid (Z_j, X_j, \phi)$ and $X_j \perp\!\!\!\perp (X_i)_{i=1}^n \mid \psi$ under $\mathcal{O}$, for all $j = n + 1, \dots$. As outlined in Section 1, the existence of such parameters is implied by the partial or unrestricted exchangeability assumptions. Given an observed realization $(w_i)_{i=1}^n$, a parametric counterpart to (3.6) can be given as

$$\begin{aligned}&\int_{\phi,\psi} \sum_x \bigl(E[Y_j \mid Z_j = 1, X_j = x, \phi; \mathcal{O}]\\ (3.7)\quad &\qquad - E[Y_k \mid Z_k = 0, X_k = x, \phi; \mathcal{O}]\bigr)\\ &\qquad \times P(X_j = x \mid \psi; \mathcal{O})\,\mathrm{d}Q\bigl(\phi, \psi \mid (w_i)_{i=1}^n; \mathcal{O}\bigr).\end{aligned}$$

It is apparent from (3.7) that drawing causal inferences is possible if the stability property of Section 3.3 applies to the pairwise comparisons

$$\begin{aligned}&E[Y_j \mid Z_j = 1, X_j = x, \phi; \mathcal{O}]\\ &\qquad - E[Y_k \mid Z_k = 0, X_k = x, \phi; \mathcal{O}],\end{aligned}$$

with ϕ parametrizing the causal effect of Z_i on Y_i when controlling for X_i as in this case, the inferences would be the same as under the experimental design. However, parametrizing causal effects directly would be reliant on statistical models, whereas the convention in causal inference literature, especially in potential outcome formulations, is to define the causal contrasts of interest first without reference to models. We address model-free definitions of causal contrasts in the present framework in Section 4.1, where parametric models may then be utilized to obtain estimators for such contrasts.

3.5 Connection to Other Latent Variable Formulations

Under the point treatment setting, the implications of the infinite extension of criterion (3.4) are equivalent to other conditions for unconfounded inferences stated in terms of conceptual latent variables representing general confounding. For instance, Definition 1 of Arjas (2012) connects unconfounded inferences to the conditional independence property $Z_i \perp\!\!\!\perp U_i \mid X_i$. This in turn directly implies that $Z_i \perp\!\!\!\perp f(z, x, U_i) \mid X_i = x$, and further the stability property similar to Section 3.3.

Although formulations in terms of latent variables need not rely on causal graphs, the absence of unmeasured confounders can be stated equivalently in terms of the backdoor criterion of Pearl (2009, p. 79); in the absence of a direct arrow $U_i \longrightarrow Z_i$ in the right-hand panel of Figure 1, X_i blocks every path between Z_i and Y_i that contains an arrow into Z_i (and is not a descendant of Z_i), which implies that X_i is sufficient to control for confounding. Alternatively, the conditional independence property $Z_i \perp\!\!\!\perp U_i \mid X_i$ can be read directly from the graph of Figure 1 using, for example, the moralization criterion of Lauritzen et al. (1989).

3.6 Connection to the Potential Outcomes Notation

Under the structural definition of the outcome, we can take the potential outcomes of individual i to be determined by $Y_i(z) = f(z, X_i, U_i)$ (cf. Pearl, 2009, p. 98), with the observed outcome given by $Y_i = Y_i(Z_i)$ (the latter is known as the *consistency* assumption, e.g., Cole and Frangakis, 2009, VanderWeele, 2009a, Pearl, 2010). As discussed in the previous two sections, the infinite extension of the symmetry property (3.4) implies that $Z_i \perp\!\!\!\perp f(z, x, U_i) \mid X_i = x$ under $\mathcal{O}$. This is equivalent to the statement

$$Y_i(z) \perp\!\!\!\perp Z_i \mid X_i, \tag{3.8}$$

which is in fact the probabilistic conditional exchangeability condition as defined by Hernán and Robins (2006, p. 579), or a consequence of the first part of the *strongly ignorable treatment assignment* condition, as defined by Rosenbaum and Rubin (1983, p. 43). We note that making statements about the joint distribution of the potential outcomes (as in strong ignorability) is not necessary for identification of causal contrasts; the weak version involving (3.8) suffices. Although Rubin (1978, p. 41) uses the term exchangeability in the usual Bayesian sense to justify an i.i.d. model construction, as far as we know, the connection between the Bayesian notion of exchangeability and the condition stated in terms of potential outcomes has not been made or studied within the framework of Rubin's causal model (as termed by Holland, 1986). In contrast, this connection is implied in Greenland and Robins (1986), Greenland, Robins and Pearl (1999) and Greenland and Robins (2009).

We note that, under the probabilistic exchangeability condition (3.4), we had that $f(z, x, U_j) \mid (Z_j = 1, Z_k = 0, X_j = X_k = x) \overset{\mathrm{d}}{=} f(z, x, U_k) \mid (Z_j = 1, Z_k = 0, X_j = X_k = x)$, that is, the remaining determinants of the outcome under the structural model have the same population distribution between the treatment groups. Requiring that these determinants also have the same empirical distribution between the groups being compared would correspond to the *deterministic exchangeability* condition laid out by Greenland and Robins (1986, p. 415). This is unnecessarily strong for unconfounded inferences; it rules out both confounding and imbalance (e.g., the chance imbalances that could arise even under complete randomization). If we could condition on all of the determinants of the outcome, the symmetry property conditional on $A_{xu} \equiv \{Z_j = 1, Z_k = 0, X_j = X_k = x, U_j = U_k = u\}$ could be written as

$$\begin{aligned}
&\Pr(Y_j = y, Y_k = y' \mid A_{xu}; \mathcal{O}) \\
&\quad = \Pr(f(1, X_j, U_j) = y, f(0, X_k, U_k) = y' \mid A_{xu}; \mathcal{O}) \\
&\quad = \Pr(f(0, X_j, U_j) = y', f(1, X_k, U_k) = y \mid A_{xu}; \mathcal{O}) \\
&\quad = \begin{cases} 1, & \text{when } (y, y') = (f(1, x, u), f(0, x, u)), \\ 0, & \text{when } (y, y') \neq (f(1, x, u), f(0, x, u)). \end{cases}
\end{aligned}$$

Thus, with this conditioning, the outcome is a deterministic function of the treatment assignment, and exchangeability applies trivially. This level of conditioning would be required for identifying individual level causal effects, which is impossible in practice (the "fundamental problem of causal inference" as discussed by Holland, 1986). The probabilistic condition is sufficient for identifying population-level effects. In the following section, we connect our concept of conditional exchangeability to Bayesian causal inference.

4. DEFINITION AND ESTIMATION OF CAUSAL CONTRASTS

4.1 Causal Contrasts Defined in Terms of Posterior Predictive Expectations

As noted by Greenland (2012), causal inference can alternatively be formulated as a prediction problem or a missing data problem; the potential outcomes notation corresponds to the latter formulation. In the Bayesian framework, a causal contrast of interest may be naturally defined in terms of posterior predictive expectations for further exchangeable individuals under the hypothetical experimental setting already introduced above. We define the causal contrast of interest under the randomized setting in terms of the limits

$$\begin{aligned}
&\lim_{n \longrightarrow \infty} E[Y_j \mid Z_j = z, (w_i)_{i=1}^n; \mathcal{E}] \\
&\quad - \lim_{n \longrightarrow \infty} E[Y_k \mid Z_k = z', (w_i)_{i=1}^n; \mathcal{E}],
\end{aligned} \tag{4.1}$$

where $j \neq k > n$ and $(w_i)_{i=1}^n$ is a hypothetical exchangeable sequence under $\mathcal{E}$. We may consider such a contrast for arbitrary settings of the treatment indicators z and z', thus mimicking the classical "intervention" formulation of the causal contrast. Note, however, that no special mathematical definitions or tools, other than those associated with fundamental exchangeability concepts, are required in this definition.

By de Finetti's representation theorem, the joint distribution of the data may be written

$$\begin{aligned}
&\Pr((W_i)_{i=1}^n; \mathcal{E}) \\
&\quad = \int_\theta \prod_{i=1}^n P(W_i \mid \theta; \mathcal{E}) \,\mathrm{d}Q(\theta; \mathcal{E}) \\
&\quad = \int_{\phi,\psi} \prod_{i=1}^n [P(Y_i \mid z_i, x_i, \phi; \mathcal{E}) \\
&\qquad \times P(X_i \mid \psi; \mathcal{E})] \mathrm{d}Q(\phi, \psi; \mathcal{E}) \\
&\qquad \times \int_\gamma \prod_{i=1}^n P(Z_i \mid \gamma; \mathcal{E}) \,\mathrm{d}Q(\gamma; \mathcal{E}),
\end{aligned} \tag{4.2}$$

where $\theta = (\phi, \gamma, \psi)$ represents a partition of the joint parameter vector corresponding to the above factorization of the joint parameter-conditional distribution of $W_i = (Y_i, Z_i, X_i)$ in the second line, provided parameter Γ is a priori independent of the parameters (Φ, Ψ) (cf. Gelman et al., 2004, pp. 354–355). Because all of these parameters are defined under $\mathcal{E}$, this independence follows from the factorization (2.1), understanding the parameters as long-run summaries of the observable sequences. Now for any $j > n$ the expectations in (4.1) may be written as

$$\begin{aligned} & E[Y_j \mid z_j, (w_i)_{i=1}^n; \mathcal{E}] \\ & \quad = \sum_{y_j, x_j} y_j \Pr(y_j, x_j \mid z_j, (w_i)_{i=1}^n; \mathcal{E}) \\ & \quad = \frac{\sum_{y_j, x_j} y_j \int_{\phi,\psi} \prod_{i \in \{1,\dots,n,j\}} L_i(\phi, \psi)\, \mathrm{d}Q(\phi, \psi; \mathcal{E})}{\sum_{y_j, x_j} \int_{\phi,\psi} \prod_{i \in \{1,\dots,n,j\}} L_i(\phi, \psi)\, \mathrm{d}Q(\phi, \psi; \mathcal{E})} \\ & \quad = \sum_{y_j, x_j} y_j \int_{\phi,\psi} L_j(\phi, \psi)\, \mathrm{d}Q(\phi, \psi \mid (w_i)_{i=1}^n; \mathcal{E}), \end{aligned} \tag{4.3}$$

where $L_i(\phi, \psi) \equiv P(y_i \mid z_i, x_i, \phi; \mathcal{E}) P(x_i \mid \psi; \mathcal{E})$. Here, the terms involving parameters Γ cancel out because $Z_i \perp\!\!\!\perp X_i$ under $\mathcal{E}$ (and $\Gamma \perp\!\!\!\perp (\Phi, \Psi)$); note that this would not hold under the observational setting $\mathcal{O}$.

If we further assume regularity conditions that allow interchanging the order of limit and integration, the limit of the above expectation becomes

$$\begin{aligned} & \lim_{n \longrightarrow \infty} E[Y_j \mid z_j, (w_i)_{i=1}^n; \mathcal{E}] \\ & \quad = \sum_{x_j} \int_{\phi,\psi} E[Y_j \mid z_j, x_j, \phi; \mathcal{E}] P(x_j \mid \psi; \mathcal{E}) \\ & \qquad \times \delta_{\phi_0}(\phi) \delta_{\psi_0}(\psi)\, \mathrm{d}\phi\, \mathrm{d}\psi \\ & \quad = \sum_{x_j} E[Y_j \mid z_j, x_j, \phi_0; \mathcal{E}] P(x_j \mid \psi_0; \mathcal{E}), \end{aligned} \tag{4.4}$$

assuming that the posterior distribution converges to a degenerate distribution at the true parameter values (ϕ_0, ψ_0) (cf. van der Vaart, 1998, p. 139). The right-hand side here corresponds to the direct standardization/backdoor formula, which was previously obtained informally as equation (3.6). Because we interpret parameters as (unknown) functions of infinite sequences of observables (following Bernardo and Smith, 1994, p. 173, and as per the definitions in Section 1), identity (4.4) motivates definition (4.1) as the causal parameter of interest, as (4.4) does not depend on the prior $Q(\phi, \psi; \mathcal{E})$.

4.2 Estimation Under the Observational Setting

Identification. To estimate the causal contrast (4.1) defined under the experimental setting $\mathcal{E}$ based on data collected under the observational setting $\mathcal{O}$, in (4.3) we have to make the substitutions $P(Y_i \mid z_i, x_i, \phi; \mathcal{E}) = P(Y_i \mid z_i, x_i, \phi; \mathcal{O})$ and $P(X_i \mid \psi; \mathcal{E}) = P(X_i \mid \psi; \mathcal{O})$; the former corresponds to the stability assumption, which in turn is implied by the infinite extension of the conditional exchangeability property (3.4). The latter can be taken to be true by definition, that is, the standard population is chosen according to the observed covariate distribution. Under these assumptions, parameters ϕ and ψ have the *same* interpretation under both settings $\mathcal{E}$ and $\mathcal{O}$. With a given observed realization $(w_i)_{i=1}^n$, this gives an *estimator* for (4.4) as

$$\begin{aligned} & \sum_{x_j} \int_{\phi,\psi} E[Y_j \mid z_j, x_j, \phi; \mathcal{O}] P(x_j \mid \psi; \mathcal{O}) \\ & \qquad \times \mathrm{d}Q(\phi, \psi \mid (w_i)_{i=1}^n; \mathcal{O}). \end{aligned} \tag{4.5}$$

We may also wish to state an identifiability condition in frequency-based terms. Because (4.5) is taken to be the estimator of parameter (4.4), it is natural to require consistency, which we have if $\lim_{n \longrightarrow \infty} \mathrm{d}Q(\phi, \psi \mid (w_i)_{i=1}^n; \mathcal{O}) = \delta_{\phi_0}(\phi) \delta_{\psi_0}(\psi)$. In other words, the inferences will be *unconfounded* if

$$\begin{aligned} & \sum_{x_j} E[Y_j \mid z_j, x_j, \phi_0; \mathcal{O}] P(x_j \mid \psi_0; \mathcal{O}) \\ & \quad = \lim_{n \longrightarrow \infty} E[Y_j \mid z_j, (w_i)_{i=1}^n; \mathcal{E}]. \end{aligned}$$

A causal contrast could be defined alternatively in terms of potential outcome variables as $E[Y_i(1)] - E[Y_i(0)]$. For unconfounded inferences, we could then require that

$$\sum_{x_i} E[Y_i \mid Z_i = z, x_i; \mathcal{O}] P(x_i; \mathcal{O}) = E[Y_i(z)], \tag{4.6}$$

which follows from (3.8) (e.g., Hernán and Robins, 2006, p. 579), and makes no explicit reference to the parametrization of the problem.

Positivity. To ensure that the conditional distributions above are well defined, we need an additional assumption known as *positivity*, that is, absolute continuity of the two measures under $\mathcal{E}$ and $\mathcal{O}$ (cf. Dawid and Didelez, 2010, p. 196), stated as $P(Z_i \mid x_i, \gamma; \mathcal{E}) \ll P(Z_i \mid x_i, \gamma; \mathcal{O})$, which is equivalent to $P(Z_i \mid x_i, \gamma; \mathcal{O}) = 0 \Rightarrow P(Z_i \mid x_i, \gamma; \mathcal{E}) = 0$ or $P(Z_i \mid x_i, \gamma; \mathcal{E}) \neq 0 \Rightarrow P(Z_i \mid x_i, \gamma; \mathcal{O}) \neq 0$. In particular, if the treatment Z_i depends deterministically on the covariates X_i, inference across the observational and experimental settings would not be possible.

Estimation of expectations (4.3) may be carried out using Monte Carlo integration by sampling from the posterior distribution of (Φ, Ψ). Because the distributions $P(Y_i \mid z_i, x_i, \phi; \mathcal{O})$ and $P(X_i \mid \psi; \mathcal{O})$ implied by the representation theorem are unknown, these have to be replaced with statistical models in practice. These models do not necessarily have to be parametric (i.e., having finite-dimensional Φ and Ψ; cf. Bernardo and Smith, 1994, p. 228), for instance, we would usually model $P(X_i \mid \psi; \mathcal{O})$ with the empirical distribution of X_i; however, in practice, the curse of dimensionality limits the use of nonparametric specifications for the outcome model, and dimension-reducing modeling assumptions will become a necessity. When finite-dimensional parametrizations are used, model misspecification becomes a potential issue. In particular, one may lose the important property of valid inferences under the null hypothesis of no treatment effect, which will be elaborated on in the following section.

5. LONGITUDINAL SETTING: EXCHANGEABILITY AND SEQUENTIAL RANDOMIZATION

Having established the Bayesian formulation of causal inference in point treatment settings, we now seek to extend this reasoning to the longitudinal case, where confounding structures may be more complex. For simplicity, we consider the two time-point case and contend that the extension to multiple time-points follows straightforwardly. Consider now the slightly more complicated setting in the DAG in the bottom panel of Figure 2, labeled by $\mathcal{O}$, adapted from Robins and Wasserman (1997), where the design variables Z_{1i} and Z_{2i} represent the treatment assignment to initiate or receive a particular dose of antiretroviral medication starting at baseline and at a subsequent reexamination, respectively, for individual i. Further, let X_i represent observed anemia status at the reexamination, and Y_i an HIV viral load outcome, measured at the end of follow-up after sufficient time has passed from the reexamination. Latent variable U_i again represents the underlying immune function of individual i, which is expected to be a determinant of both X_i and Y_i. Here, X_i being influenced by earlier treatment introduces treatment-confounder feedback (Robins, Hernán and Brumback, 2000, p. 550), which makes the judgment of exchangeability somewhat more involved. For the causal exchangeability considerations, we take the intermediate variable and outcome to be determined by structural models $X_i = g(Z_{1i}, U_i)$ and $Y_i = f(Z_{1i}, Z_{2i}, X_i, U_i)$.

The principal source of difficulty is represented by the latent variable U_i. To consider its implications for inference, we first define the causal contrast of interest in terms of a randomized setting labeled by $\mathcal{E}$ depicted in the top panel of Figure 2. We may now define the causal contrast of interest as

$$\lim_{n\longrightarrow\infty} E\big[Y_j \mid Z_{1j} = z_{1j}, Z_{2j} = z_{2j}, (w_i)_{i=1}^n; \mathcal{E}\big] - \lim_{n\longrightarrow\infty} E\big[Y_k \mid Z_{1k} = z_{1k}, Z_{2k} = z_{2k}, (w_i)_{i=1}^n; \mathcal{E}\big],$$

where $j \neq k > n$. The expectations here can be represented alternatively as

$$\lim_{n\longrightarrow\infty} E\big[Y_i \mid z_{1i}, z_{2i}, (w_i)_{i=1}^n; \mathcal{E}\big] = E[Y_i \mid z_{1i}, z_{2i}, \varphi_0; \mathcal{E}] \tag{5.1}$$

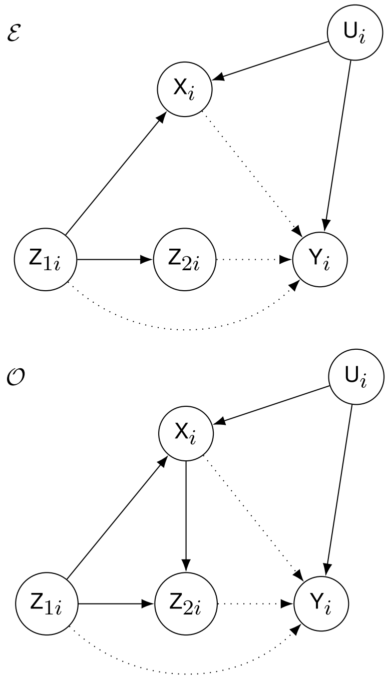

FIG. 2. *Top panel*: *DAG depicting the randomized longitudinal setting labeled by $\mathcal{E}$. The dashed arrows are absent under the null hypothesis of no treatment effect in the presence of treatment-confounder feedback. Note that the null hypothesis also holds under an alternative DAG, where the arrow $Z_{1i} \longrightarrow X_i$ is omitted, and the dotted arrow $X_i \longrightarrow Y_i$ may be present. Bottom panel*: *DAG depicting the observational longitudinal setting labeled by $\mathcal{O}$. This DAG differs from that on the top panel by the arrow $X_i \longrightarrow Z_{2i}$.*

or

$$\begin{aligned} &\lim_{n\longrightarrow\infty} E\big[Y_i \mid z_{1i}, z_{2i}, (w_i)_{i=1}^n; \mathcal{E}\big] \\ &\quad = \sum_{x_i} E\big[Y_i \mid z_{1i}, z_{2i}, x_i, \phi_0^*; \mathcal{E}\big] \\ &\qquad \times P\big(x_i \mid z_{1i}, \psi_0^*; \mathcal{E}\big) \end{aligned} \tag{5.2}$$

or finally

$$\begin{aligned} &\lim_{n\longrightarrow\infty} E\big[Y_i \mid z_{1i}, z_{2i}, (w_i)_{i=1}^n; \mathcal{E}\big] \\ &\quad = \sum_{x_i} \int_{u_i} E\big[Y_i \mid z_{1i}, z_{2i}, x_i, u_i, \phi_0^\dagger; \mathcal{E}\big] \\ &\qquad \times P\big(x_i \mid z_{1i}, u_i, \psi_0^\dagger; \mathcal{E}\big) P\big(\mathrm{d}u_i \mid \eta_0^\dagger; \mathcal{E}\big). \end{aligned} \tag{5.3}$$

Note the different parameters φ, (ϕ^*, ψ^*) and $(\phi^\dagger, \psi^\dagger, \eta^\dagger)$ in the three representations. The parametrization in (5.3) corresponds to the data generating mechanism, the parameters of which are determined by the representation for infinitely exchangeable random vectors $(Y_i, Z_{1i}, Z_{2i}, X_i, U_i)$, whereas the parameters that appear in (5.1) and (5.2) are consequences of the joint model obtained by marginalization.

As was done in Section 3.2, we consider for simplicity binary or dichotomized treatments and consider the comparability of groups selected to have a given treatment assignment configuration. The exchangeability with respect to the intermediate variable X_i can be established as before. For the outcome Y_i, we consider exchangeability separately at the time of each treatment. The groups being compared have the treatment assignments $(Z_{1i} = 1, Z_{2i} = 1)$, $(Z_{1i} = 1, Z_{2i} = 0)$, $(Z_{1i} = 0, Z_{2i} = 1)$ and $(Z_{1i} = 0, Z_{2i} = 0)$. We note that the parameters φ in the outcome model $P(Y_i \mid z_{1i}, z_{2i}, \varphi; \mathcal{E})$ corresponding to parametrization (5.1) would not be estimatable under the observational setting $\mathcal{O}$. For instance, at the second time point, the outcomes of individuals j and k with opposite treatment assignments would not be exchangeable (those assigned to treatment at the second interval are likely to have better underlying immune function status than those not assigned to treatment, with the second assignment depending on X_i), that is, we do not have that

$$\begin{aligned}
&\Pr(Y_j = y, Y_k = y' \mid A; \mathcal{O})\\
&\quad = \Pr\bigl(f(1, 1, g(1, U_j), U_j) = y,\\
&\qquad f(1, 0, g(1, U_k), U_k) = y' \mid A; \mathcal{O}\bigr)\\
&\quad = \Pr\bigl(f(1, 1, X_j, U_j) = y,\\
&\qquad f(1, 0, X_k, U_k) = y' \mid A; \mathcal{O}\bigr)\\
&\quad = \Pr\bigl(f(1, 0, X_j, U_j) = y',\\
&\qquad f(1, 1, X_k, U_k) = y \mid A; \mathcal{O}\bigr),
\end{aligned}$$

where $A \equiv \{Z_{1j} = Z_{1k} = 1, Z_{2j} = 1, Z_{2k} = 0\}$.

Instead, we can adopt parametrization (5.2) and model the conditional distributions $P(Y_i \mid z_{1i}, z_{2i}, x_i, \phi^*; \mathcal{O})$ and $P(X_i \mid z_{1i}, \psi^*; \mathcal{E})$. Now based on Figure 2 we have that $Z_{1i} \perp\!\!\!\perp U_i$ and $Z_{2i} \perp\!\!\!\perp U_i \mid (Z_{1i}, X_i)$ under $\mathcal{O}$, which together imply the sequential randomization condition discussed by, for example, Dawid and Didelez (2010, p. 200), or stability $P(Y_i \mid Z_{1i}, Z_{2i}, X_i; \mathcal{E}) = P(Y_i \mid Z_{1i}, Z_{2i}, X_i; \mathcal{O})$ and $P(X_i \mid Z_{1i}; \mathcal{E}) = P(X_i \mid Z_{1i}; \mathcal{O})$. Stability would be sufficient to ensure nonparametric identification of the marginal causal contrast because

$$\begin{aligned}
&P(Y_i \mid z_{1i}, z_{2i}; \mathcal{E})\\
&\quad = \sum_{x_i} \int_{u_i} \frac{P(Y_i, Z_{1i} = z_{1i}, Z_{2i} = z_{2i}, x_i, \mathrm{d}u_i; \mathcal{E})}{P(Z_{1i} = z_{1i}, Z_{2i} = z_{2i}; \mathcal{E})}\\
&\quad = \sum_{x_i} \int_{u_i} P(Y_i \mid z_{1i}, z_{2i}, x_i, u_i; \mathcal{E}) P(x_i \mid z_{1i}, u_i; \mathcal{E})\\
&\qquad \times P(\mathrm{d}u_i \mid z_{1i}; \mathcal{E})\\
&\quad = \sum_{x_i} \int_{u_i} P(Y_i \mid z_{1i}, z_{2i}, x_i, u_i; \mathcal{E}) P(x_i \mid z_{1i}; \mathcal{E})\\
&\qquad \times P(\mathrm{d}u_i \mid z_{1i}, z_{2i}, x_i; \mathcal{E})\\
&\quad = \sum_{x_i} \int_{u_i} P(Y_i, \mathrm{d}u_i \mid z_{1i}, z_{2i}, x_i; \mathcal{E}) P(x_i \mid z_{1i}; \mathcal{E})\\
&\quad = \sum_{x_i} P(Y_i \mid z_{1i}, z_{2i}, x_i; \mathcal{E}) P(x_i \mid z_{1i}; \mathcal{E})\\
&\quad = \sum_{x_i} P(Y_i \mid z_{1i}, z_{2i}, x_i; \mathcal{O}) P(x_i \mid z_{1i}; \mathcal{O}).
\end{aligned}$$

However, under the longitudinal setting introducing stratification by X_i does not restore the conditional exchangeability of all the groups being compared. We now do have exchangeability between individuals j and k with opposing treatment assignments at the second time point, that is,

$$\begin{aligned}
&\Pr(Y_j = y, Y_k = y' \mid A_x; \mathcal{O})\\
&\quad = \Pr\bigl(f(1, 1, x, U_j) = y,\\
&\qquad f(1, 0, x, U_k) = y' \mid A_x; \mathcal{O}\bigr)\\
&\quad = \Pr\bigl(f(1, 0, x, U_j) = y',\\
&\qquad f(1, 1, x, U_k) = y \mid A_x; \mathcal{O}\bigr),
\end{aligned} \tag{5.4}$$

where $A_x \equiv \{Z_{1j} = Z_{1k} = 1, Z_{2j} = 1, Z_{2k} = 0, X_j = X_k = x\}$. However, when comparing individuals with opposite treatment assignments at the first time point, the conditional exchangeability condition

$$\begin{aligned}
&\Pr(Y_j = y, Y_k = y' \mid A_x; \mathcal{O})\\
&\quad = \Pr\bigl(f(1, 0, x, U_j) = y,\\
&\qquad f(0, 0, x, U_k) = y' \mid A_x; \mathcal{O}\bigr)\\
&\quad = \Pr\bigl(f(0, 0, x, U_j) = y',\\
&\qquad f(1, 0, x, U_k) = y \mid A_x; \mathcal{O}\bigr),
\end{aligned} \tag{5.5}$$

where $A_x \equiv \{Z_{1j} = 1, Z_{1k} = 0, Z_{2j} = Z_{2k} = 0, X_j = X_k = x\}$, does *not* hold because the prior information we have on the relationships between the variables indicates, for example, that those without anemia and assigned to treatment at the first interval are likely to have better immune function status than those without anemia and no treatment at the first interval because initiation of the treatment is in itself a cause of anemia. This would be the case also if the groups being compared had been formed under the completely randomized setting $\mathcal{E}$, even though the groups would be exchangeable without the stratification. In the causal inference literature, this phenomenon has been called *collider stratification bias* (e.g., Greenland, 2003), *Berkson's bias* or merely *selection bias*; as demonstrated, it can equally well be understood as lack of conditional exchangeability of the groups being compared in terms of their pretreatment characteristics. Exchangeability does hold matching on the initial treatment assignment Z_{i1}, but this would not allow estimation of the effect of Z_{i1}. The nonexchangeability of the groups not matched with respect to the initial treatment assignment is illustrated in the numerical example presented in Supplementary Appendix B.

The lack of conditional exchangeability corresponding to (5.5) implies that the parameters ϕ^* in the conditional probability model $P(Y_i \mid z_{1i}, z_{2i}, x_i, \phi^*; \mathcal{O})$ characterizing the association between Y_i and Z_{2i} would not have a causal interpretation, and thus a modeling strategy based on finite-dimensional parametrization (ϕ^*, ψ^*) might not be successful; without an appropriate parametrization of the problem, we may lose the important property of valid inferences under the null hypothesis of no treatment effect, which gives rise to the so-called *null paradox* (e.g., Robins and Wasserman, 1997, pp. 411–412, Vansteelandt, Bekaert and Claeskens, 2012, p. 11; Dawid and Didelez, 2010, p. 224).

The conditional exchangeability condition (5.5) relates to the stronger conditional independence condition $(Z_{1i}, Z_{2i}) \perp\!\!\!\perp U_i \mid X_i$ required for identification of controlled direct effects (e.g., Robins and Greenland, 1992, VanderWeele, 2009b). This does not hold under the setting of Figure 2, but exchangeability could be restored by introducing further conditioning on U_i, which implies that (5.3) would be the correct causal parametrization. However, because U_i is unobserved, the use of such parametrization in practice would introduce new identifiability problems. The null-robust reparametrization of the problem, as suggested by Robins and Wasserman (1997, pp. 415–416) might be one way to proceed.

Regardless of the issues related to finite-dimensional parametrizations, we note that a connection between conditional exchangeability statements and the stability property is still preserved in the longitudinal setting. As we have noted above, sequential randomization is sufficient for stability, and assuming conditional exchangeability under permutations of both treatment assignments Z_{1i} and Z_{2i} is unnecessarily strong for nonparametric identifiability of the problem. If we assume the infinite extension of exchangeability property (5.4) with respect to permutations Z_{2i} at fixed levels of Z_{1i}, we note that at the second time point Z_{1i} has the same role as the observed confounders X_i. We can then use the same arguments as in Sections 3.1 and 3.2 to find that $Z_{2i} \perp\!\!\!\perp f(z_1, z_2, x, U_i) \mid (Z_{1i} = z_1, X_i = x)$ under both $\mathcal{E}$ and $\mathcal{O}$. This corresponds to the second condition of sequential randomization and can be used to further obtain $\Pr(Y_i = y \mid Z_{i1} = z_1, Z_{i2} = z_2, X_i = x; \mathcal{E}) = \Pr(f(z_1, z_2, x, U_i) = y \mid Z_{i1} = z_1, Z_{i2} = z_2, X_i = x; \mathcal{E}) = \Pr(f(z_1, z_2, x, U_i) = y \mid Z_{i1} = z_1, X_i = x; \mathcal{E})$. Here,

$$\Pr\big(f(z_1, z_2, x, U_i) = y \mid Z_{i1} = z_1, X_i = x; \mathcal{E}\big) = \frac{\Pr(f(z_1, z_2, x, U_i) = y, g(z_1, U_i) = x \mid Z_{i1} = z_1; \mathcal{E})}{\Pr(g(z_1, U_i) = x \mid Z_{i1} = z_1; \mathcal{E})}.$$

Thus, if we have $(f(z_1, z_2, g(z_1, U_i), U_i), g(z_1, U_i)) \perp\!\!\!\perp Z_{1i}$ (which in turn implies that $g(z_1, U_i) \perp\!\!\!\perp Z_{1i}$), under the usual assumption that the distribution of the baseline characteristics is the same under $\mathcal{E}$ and $\mathcal{O}$, we can get that

$$\begin{aligned}
&\Pr\big(f(z_1, z_2, x, U_i) = y \mid Z_{i1} = z_1, X_i = x; \mathcal{E}\big)\\
&\quad = \Pr\big(f(z_1, z_2, x, U_i) \mid Z_{i1} = z_1, X_i = x; \mathcal{O}\big)\\
&\quad = \Pr\big(f(z_1, z_2, x, U_i) \mid Z_{i1} = z_1, Z_{i2} = z_2, X_i = x; \mathcal{O}\big)\\
&\quad = \Pr(Y_i \mid Z_{i1} = z_1, Z_{i2} = z_2, X_i = x; \mathcal{O}).
\end{aligned}$$

Using similar arguments as before, these properties also apply in the i.i.d. distribution, implying the stability property for the outcome distribution. The first sequential randomization condition $Z_{1i} \perp\!\!\!\perp U_i$ would be sufficient for the required independence, but it can also be obtained from the infinite joint exchangeability property for sequences of $f(z_1, z_2, g(z_1, U_i), U_i)$ and $g(z_1, U_i)$ conditional on Z_{1i}. Thus, we contend that while obtaining identifying conditions for causal effects based on conditional exchangeability statements is more cumbersome in the presence of treatment-confounder feedback, it appears to be possible. We also note that the required identifying conditions correspond to $(Y_i(z_1, z_2), X_i(z_1)) \perp\!\!\!\perp Z_{1i}$ and $Y_i(z_1, z_2) \perp\!\!\!\perp Z_{2i} \mid (Z_{1i}, X_i)$ expressed in terms of potential outcome variables if we take $Y_i(z_1, z_2) = f(z_1, z_2, g(z_1, U_i), U_i)$ and $X_i(z_1) = g(z_1, U_i)$, that is, the treatment assignments are independent of future potential outcomes and intermediate variables conditional on observed past (e.g., Chakraborty and Murphy, 2014).

6. DISCUSSION

We have demonstrated that the notion of exchangeability as a probabilistic symmetry property can indeed serve as as a basis of a causal model, as was originally suggested by Lindley and Novick (1981). That exchangeability can be formulated as an ignorability assumption, and that marginal causal contrasts can be naturally defined in terms of limits of posterior predictive expectations for further, yet unobserved, exchangeable individuals, has not been appreciated in the causal inference literature. We do not claim that the interpretation of exchangeability as a causal model would have important practical advantages over alternative causal models; the preference for a particular causal model as the notational system is largely a matter of taste and convention. In particular, the identifying conditions required for inferences were equivalent to corresponding conditions stated in terms of potential outcomes. However, the proposed framework links causality more closely to model parameters and does enable a more natural incorporation of causal reasoning into the fully probabilistic Bayesian framework, in the sense that no concepts external to de Finetti's system are necessary.

We demonstrated a connection between conditional exchangeability statements and causal interpretation of parameters in statistical models. However, in the longitudinal setting of Section 5, the connection between the conditional exchangeability properties corresponding to the model components and identifying conditions for marginal causal contrasts defined without reference to statistical models becomes less direct. In particular, in situations where conditioning on intermediate variables opens backdoor paths between treatments and the outcome, component models may not be interpretable, while the marginal causal effects may still be identifiable. Alternative inference methods exist that can identify the causal contrast under the sequential randomization condition and with fewer parametric modeling assumptions; consider, for example, marginal structural models estimated using inverse probability of treatment weighting (Robins, Hernán and Brumback, 2000, Hernán, Brumback and Robins, 2001). Nonetheless, exchangeability judgments may warn us of a situation where null paradox-type model misspecification issues are likely to arise. Proper understanding of the problem and the possible solutions are especially important given the recent renewed interest in the parametric g-computation formula (e.g., Taubman et al., 2009, Westreich et al., 2012, Keil et al., 2014, Jain et al., 2016, Bijlsma et al., 2017, Neophytou et al., 2019, Shahn et al., 2019). The issues related to finite-dimensional parametrizations also motivate further research into semiparametric Bayesian inference procedures, which would allow direct parametrization of marginal causal effects while avoiding specifications of some of the likelihood components (cf. Saarela et al., 2015, Saarela, Belzile and Stephens, 2016).

Throughout, we assumed a functional relationship between the outcome and its determinants, with the function $f(z, X_i, U_i)$ understood as the equivalent of the potential outcome $Y(z)$. This notation allows us to decouple the observed, potentially informative, treatment assignment from the intervention in the exchangeability judgments when considering switching the treatment of the units. The assumed deterministic relationship may not be a serious limitation, as U_i could always be thought to include the remaining (unobserved) determinants of the outcome. However, a reviewer points out that the present framework could potentially be modified to allow for stochastic dependency of Y on (Z, X, U) by introducing separate notation for the intended/assigned treatment Z and intervention to administer treatment $\widehat{Z}$. One could then consider exchangeability statements of the type

$$\begin{aligned}&\Pr(Y_j = y; Y_k = y' \mid Z_j = z, Z_k = z'; \widehat{Z}_j = z, Z_k = z')\\&\quad = \Pr(Y_j = y'; Y_k = y \mid Z_j = z, Z_k = z';\\&\qquad \widehat{Z}_j = z', Z_k = z),\end{aligned}$$

conditional on both the assignment and the intervention (which may be different). This has a similar interpretation as (3.1) but does not require introducing the functional relationship for the outcome. We leave it as further work to study whether the presented framework can be adapted accordingly to obtain the same results.

APPENDIX: LIKELIHOOD CONSTRUCTION UNDER DICHOTOMOUS OUTCOMES

Suppose that $Y_i \in \{0, 1\}$ is an outcome event indicator or dichotomized continuous or count outcome and the subsequences of these indicator variables for treated units $Z_i = 1$ and untreated units $Z_i = 0$ are separately infinitely exchangeable, that is, we have partial exchangeability

$$\begin{aligned}&\Pr\Biggl(\bigcap_{i:z_i=1} (Y_i = y_i), \bigcap_{i:z_i=0} (Y_i = y_i) \Bigm| \bigcap_{i=1}^{n} (Z_i = z_i); \mathcal{E}\Biggr)\\&\quad = \Pr\Biggl(\bigcap_{i:z_i=1} (Y_i = y_{\rho_1(i)}),\\&\qquad \bigcap_{i:z_i=0} (Y_i = y_{\rho_0(i)}) \Bigm| \bigcap_{i=1}^{n} (Z_i = z_i); \mathcal{E}\Biggr)\end{aligned}$$

for any permutations ρ_1 and ρ_0 of the subsequences. If in addition, we assume that the treated event count and untreated event count are sufficient statistics, by Proposition 4.18 of Bernardo and Smith (1994), for each pair of treated and untreated units we have that

$$\begin{aligned}&\Pr(Y_j = y, Y_k = y' \mid Z_j = 1, Z_k = 0; \mathcal{E})\\&\quad = \int_{[0,1]^2} \Pr(Y_j = y, Y_k = y' \mid Z_j = 1, Z_k = 0,\\&\qquad \phi_1, \phi_0; \mathcal{E}) \, \mathrm{d}Q(\phi; \mathcal{E})\\&\quad = \int_{[0,1]^2} \phi_1^{y} (1 - \phi_1)^{1-y} \phi_0^{y'} (1 - \phi_0)^{1-y'} \, \mathrm{d}Q(\phi; \mathcal{E}),\end{aligned}$$

where $\phi = (\phi_0, \phi_1)$,

$$\begin{aligned}&Q(\phi; \mathcal{E})\\&\quad = \lim_{n \longrightarrow \infty} \Pr\Biggl(\frac{\sum_{i=1}^{n} z_i Y_i}{\sum_{i=1}^{n} z_i} \le \phi_1,\\&\qquad \frac{\sum_{i=1}^{n} (1 - z_i) Y_i}{\sum_{i=1}^{n} (1 - z_i)} \le \phi_0 \Biggm| \bigcap_{i=1}^{n} (Z_i = z_i)\Biggr)\\&\quad = \lim_{n \longrightarrow \infty} \Pr\Biggl(\frac{\sum_{i=1}^{n} z_i f(1, X_i, U_i)}{\sum_{i=1}^{n} z_i} \le \phi_1,\\&\qquad \frac{\sum_{i=1}^{n} (1 - z_i) f(0, X_i, U_i)}{\sum_{i=1}^{n} (1 - z_i)} \le \phi_0 \Biggm| \bigcap_{i=1}^{n} (Z_i = z_i)\Biggr),\end{aligned}$$

and

$$\begin{aligned}\phi_1 &= \lim_{n \longrightarrow \infty} \frac{\sum_{i=1}^{n} z_i Y_i}{\sum_{i=1}^{n} z_i}\\&= \lim_{n \longrightarrow \infty} \frac{\sum_{i=1}^{n} z_i f(1, X_i, U_i)}{\sum_{i=1}^{n} z_i}\end{aligned}$$

and

$$\phi_0 = \lim_{n \longrightarrow \infty} \frac{\sum_{i=1}^n (1 - z_i) Y_i}{\sum_{i=1}^n (1 - z_i)}$$
$$= \lim_{n \longrightarrow \infty} \frac{\sum_{i=1}^n (1 - z_i) f(0, X_i, U_i)}{\sum_{i=1}^n (1 - z_i)}.$$

As noted before, for contrasts of the treated and untreated outcome event frequencies such as $\phi_1 - \phi_0$ or ϕ_1/ϕ_0 to have any causal interpretation, we need further assumptions in addition to the partial exchangeability. From the general exchangeability, it follows that

$$\begin{aligned}
&\Pr(Y_j = y, Y_k = y' \mid Z_j = 1, Z_k = 0; \mathcal{E}) \\
&\quad = \frac{\Pr(Y_j = y, Y_k = y', Z_j = 1, Z_k = 0; \mathcal{E})}{\Pr(Z_j = 1, Z_k = 0; \mathcal{E})} \\
&\quad = \frac{\Pr(Y_j = y', Y_k = y, Z_j = 0, Z_k = 1; \mathcal{E})}{\Pr(Z_j = 0, Z_k = 1; \mathcal{E})} \\
&\quad = \Pr(Y_j = y', Y_k = y \mid Z_j = 0, Z_k = 1; \mathcal{E}).
\end{aligned}$$

Further, by the structural model and (3.1), we have that

$$\begin{aligned}
&\Pr(Y_j = y', Y_k = y \mid Z_j = 0, Z_k = 1; \mathcal{E}) \\
&\quad = \Pr(f(1, X_k, U_k) = y, \\
&\qquad f(0, X_j, U_j) = y' \mid Z_j = 0, Z_k = 1; \mathcal{E}) \\
&\quad = \Pr(f(1, X_j, U_j) = y, \\
&\qquad f(0, X_k, U_k) = y' \mid Z_j = 0, Z_k = 1; \mathcal{E}),
\end{aligned}$$

so that

$$\begin{aligned}
&\Pr(f(1, X_j, U_j) = y, \\
&\qquad f(0, X_k, U_k) = y' \mid Z_j = 1, Z_k = 0; \mathcal{E}) \\
&\quad = \Pr(f(1, X_j, U_j) = y, \\
&\qquad f(0, X_k, U_k) = y' \mid Z_j = 0, Z_k = 1; \mathcal{E}).
\end{aligned}$$

This indicates that under the added conditional exchangeability assumption, the joint distribution of $f(1, X_j, U_j)$ and $f(0, X_k, U_k)$ does not depend on which one of j and k was actually assigned the treatment. More generally, we have that

$$\begin{aligned}
&\Pr\Bigl(\bigcap_{i=1}^n (f(z_i, X_i, U_i) = y_i) \mid \bigcap_{i=1}^n (Z_i = z_i); \mathcal{E}\Bigr) \\
&\quad = \Pr\Bigl(\bigcap_{i=1}^n (f(z_i, X_i, U_i) = y_i) \mid \bigcap_{i=1}^n (Z_{\rho(i)} = z_i); \mathcal{E}\Bigr).
\end{aligned}$$

This indicates that the limiting distribution $Q(\phi; \mathcal{E})$ and the limits ϕ_1 and ϕ_0 do not depend on the actual treatment assignment. We also note that from (3.2) it follows that

$$\begin{aligned}
&\Pr\Bigl(\bigcap_{i:z_i=1} (f(1, X_i, U_i) = y_i), \\
&\qquad \bigcap_{i:z_i=0} (f(0, X_i, U_i) = y_i) \mid \bigcap_{i=1}^n (Z_i = z_i); \mathcal{E}\Bigr) \\
&\quad = \Pr\Bigl(\bigcap_{i:z_i=1} (f(1, X_i, U_i) = y_{\rho_1(i)}), \\
&\qquad \bigcap_{i:z_i=0} (f(0, X_i, U_i) = y_{\rho_0(i)}) \mid \bigcap_{i=1}^n (Z_i = z_i); \mathcal{E}\Bigr)
\end{aligned}$$

for any permutations ρ_1 and ρ_0 of the treated and untreated subsequences. This means that partial exchangeability still applies to the sequences of under treatment random variables $f(1, X_i, U_i)$ and without treatment random variables $f(0, X_i, U_i)$, enabling application of Proposition 4.18 of Bernardo and Smith (1994) directly to the joint distribution of these. Thus, we conclude that the likelihood is given by

$$\begin{aligned}
&\Pr(Y_j = y, Y_k = y' \mid Z_j = 1, Z_k = 0, \phi_1, \phi_0; \mathcal{E}) \\
&\quad = \Pr(f(1, X_j, U_j) = y, \\
&\qquad f(0, X_k, U_k) = y' \mid Z_j = 1, Z_k = 0, \phi_1, \phi_0; \mathcal{E}) \\
&\quad = \Pr(f(1, X_k, U_k) = y, \\
&\qquad f(0, X_j, U_j) = y' \mid Z_j = 1, Z_k = 0, \phi_1, \phi_0; \mathcal{E}),
\end{aligned}$$

so symmetry property (3.1) is preserved conditional on the parameters.

ACKNOWLEDGMENTS

The authors would like to thank an anonymous reviewer and the Editor for their thorough and constructive feedback that helped us to sharpen and solidify our argument. The first author would like to thank Professor Dennis Lindley (1923–2013) for the inspiration for this work and for hosting him and Professor Elja Arjas in Minehead in April 2012.

SUPPLEMENTARY MATERIAL

Supplement to "The Role of Exchangeability in Causal Inference" (DOI: 10.1214/22-STS879SUPP; .pdf). Supplementary Appendix A: Includes Bayesian DAGs. Supplementary Appendix B: Includes numerical examples.

REFERENCES

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

ARJAS, E. and PARNER, J. (2004). Causal reasoning from longitudinal data. *Scand. J. Stat.* **31** 171–187. MR2066247 https://doi.org/10.1111/j.1467-9469.2004.02-134.x

BAKER, S. G. (2013). Causal inference, probability theory, and graphical insights. *Stat. Med.* **32** 4319–4330. MR3118357 https://doi.org/10.1002/sim.5828

BERNARDO, J.-M. and SMITH, A. F. M. (1994). *Bayesian Theory*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, Chichester. MR1274699 https://doi.org/10.1002/9780470316870

BIJLSMA, M. J., TARKIAINEN, L., MYRSKYLÄ, M. and MARTIKAINEN, P. (2017). Unemployment and subsequent depression: A mediation analysis using the parametric G-formula. *Soc*. *Sci*. *Med*. **194** 142–150.

BÜHLMANN, P. (2020). Invariance, causality and robustness: 2018 Neyman Lecture. *Statist*. *Sci*. **35** 404–426. MR4148216 https://doi.org/10.1214/19-STS721

CHAKRABORTY, B. and MURPHY, S. A. (2014). Dynamic treatment regimes. *Annu*. *Rev*. *Stat*. *Appl*. **1** 447–464.

CHIB, S. (2007). Analysis of treatment response data without the joint distribution of potential outcomes. *J*. *Econometrics* **140** 401–412. MR2408912 https://doi.org/10.1016/j.jeconom.2006.07.009

COLE, S. R. and FRANGAKIS, C. E. (2009). The consistency statement in causal inference: A definition or an assumption? *Epidemiology* **20** 3–5.

COMMENGES, D. (2019). Causality without potential outcomes and the dynamic approach. Preprint. Available at arXiv:1905.01195.

COMMENGES, D. and GÉGOUT-PETIT, A. (2015). The stochastic system approach for estimating dynamic treatments effect. *Lifetime Data Anal*. **21** 561–578. MR3397506 https://doi.org/10.1007/s10985-015-9322-3

DAWID, A. P. (2000). Causal inference without counterfactuals. *J*. *Amer*. *Statist*. *Assoc*. **95** 407–448. MR1803167 https://doi.org/10.2307/2669377

DAWID, P. (2021). Decision-theoretic foundations for statistical causality. *J*. *Causal Inference* **9** 39–77. MR4289525 https://doi.org/10.1515/jci-2020-0008

DAWID, A. P. and DIDELEZ, V. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Stat*. *Surv*. **4** 184–231. MR2740837 https://doi.org/10.1214/10-SS081

DAWID, A. P., MUSIO, M. and FIENBERG, S. E. (2016). From statistical evidence to evidence of causality. *Bayesian Anal*. **11** 725–752. MR3498044 https://doi.org/10.1214/15-BA968

DE FINETTI, B. (1929). Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici*: *Bologna del* 3 *al* 10 *de Settembre di* 1928 179–190.

DE FINETTI, B. (1938). Sur la condition d'équivalence partielle. *Actual*. *Sci*. *Ind*. **739**. Translated In: Studies in Inductive and Probability, II. Jeffrey, R. (ed.) University of California Press: Berkeley 1980.

DIACONIS, P. (1988). Recent progress on de Finetti's notions of exchangeability. In *Bayesian Statistics*, 3 (*Valencia*, 1987). *Oxford Sci*. *Publ*. 111–125. Oxford Univ. Press, New York. MR1008047

FERREIRA, J. A. (2015). Some models and methods for the analysis of observational data. *Stat*. *Surv*. **9** 106–208. MR3396384 https://doi.org/10.1214/15-SS110

FERREIRA, J. A. (2019). Causality from the point of view of statistics. Preprint. Available at arXiv:1908.07301.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. *Texts in Statistical Science Series*. CRC Press/CRC, Boca Raton, FL. MR2027492

GREENLAND, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology* **14** 300–306.

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

GREENLAND, S. and ROBINS, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *Int*. *J*. *Epidemiol*. **15** 413–419. https://doi.org/10.1093/ije/15.3.413

GREENLAND, S. and ROBINS, J. M. (2009). Identifiability, exchangeability, and epidemiological confounding revisited. *Epidemiol*. *Perspect*. *Innov*. **6**. https://doi.org/10.1186/1742-5573-6-4.

GREENLAND, S., ROBINS, J. M. and PEARL, J. (1999). Confounding and collapsibility in causal inference. *Statist*. *Sci*. **14** 29–46.

HERNÁN, M. A., BRUMBACK, B. and ROBINS, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J*. *Amer*. *Statist*. *Assoc*. **96** 440–448. MR1939347 https://doi.org/10.1198/016214501753168154

HERNÁN, M. A. and ROBINS, J. M. (2006). Estimating causal effects from epidemiological data. *J*. *Epidemiol*. *Community Health* **60** 578–586.

HEWITT, E. and SAVAGE, L. J. (1955). Symmetric measures on Cartesian products. *Trans*. *Amer*. *Math*. *Soc*. **80** 470–501. MR0076206 https://doi.org/10.2307/1992999

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

JAIN, P., DANAEI, G., ROBINS, J. M., MANSON, J. E. and HERNÁN, M. A. (2016). Smoking cessation and long-term weight gain in the Framingham Heart Study: An application of the parametric g-formula for a continuous outcome. *Eur*. *J*. *Epidemiol*. **31** 1223–1229. https://doi.org/10.1007/s10654-016-0200-4

KEIDING, N. and CLAYTON, D. (2014). Standardization and control for confounding in observational studies: A historical perspective. *Statist*. *Sci*. **29** 529–558. MR3300358 https://doi.org/10.1214/13-STS453

KEIL, A. P., EDWARDS, J. K., RICHARDSON, D. R., NAIMI, A. I. and COLE, S. R. (2014). The parametric g-formula for time-to-event data: Towards intuition with a worked example. *Epidemiology* **25** 889.

LAURITZEN, S. L., ANDERSEN, A. H., EDWARDS, D., JÖRESKOG, K. G. and JOHANSEN, S. (1989). Mixed graphical association models [with discussion and rejoinder]. *Scand*. *J*. *Stat*. **16** 273–306.

LINDLEY, D. V. (2002). Seeing and doing: The concept of causation. *Int*. *Stat*. *Rev*. **70** 191–214.

LINDLEY, D. V. and NOVICK, M. R. (1981). The role of exchangeability in inference. *Ann*. *Statist*. **9** 45–58. MR0600531

NEOPHYTOU, A. M., COSTELLO, S., PICCIOTTO, S., BROWN, D. M., ATTFIELD, M. D., BLAIR, A., LUBIN, J. H., STEWART, P. A., VERMEULEN, R. et al. (2019). Diesel exhaust, respirable dust, and ischemic heart disease: An application of the parametric g-formula. *Epidemiology* **30** 177–185.

PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

PEARL, J. (2010). On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? *Epidemiology* **21** 872–875. https://doi.org/10.1097/EDE.0b013e3181f5d3fd

ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155. https://doi.org/10.1097/00001648-199203000-00013

ROBINS, J. M., HERNÁN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

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

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

RØYSLAND, K. (2011). A martingale approach to continuous-time marginal structural models. *Bernoulli* **17** 895–915. MR2817610 https://doi.org/10.3150/10-BEJ303

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

SAARELA, O., BELZILE, L. R. and STEPHENS, D. A. (2016). A Bayesian view of doubly robust causal inference. *Biometrika* **103** 667–681. MR3551791 https://doi.org/10.1093/biomet/asw025

SAARELA, O., STEPHENS, D. A and MOODIE, E. E (2023). Supplement to "The Role of Exchangeability in Causal Inference." https://doi.org/10.1214/22-STS879SUPP

SAARELA, O., STEPHENS, D. A., MOODIE, E. E. M. and KLEIN, M. B. (2015). On Bayesian estimation of marginal structural models. *Biometrics* **71** 279–288. MR3366229 https://doi.org/10.1111/biom.12269

SHAHN, Z., LI, Y., SUN, Z., MOHAN, A., SAMPAIO, C. and HU, J. (2019). G-computation and hierarchical models for estimating multiple causal effects from observational disease registries with irregular visits. *AMIA Joint Summits on Translational Science Proceedings* **2019** 789–798.

TAUBMAN, S. L., ROBINS, J. M., MITTLEMAN, M. A. and HERNÁN, M. A. (2009). Intervening on risk factors for coronary heart disease: An application of the parametric g-formula. *Int. J. Epidemiol.* **38** 1599–1611. https://doi.org/10.1093/ije/dyp192

VANDERWEELE, T. J. (2009a). Concerning the consistency assumption in causal inference. *Epidemiology* **20** 880–883. https://doi.org/10.1097/EDE.0b013e3181bd5638

VANDERWEELE, T. J. (2009b). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* **20** 18–26. https://doi.org/10.1097/EDE.0b013e31818f69ce

VANSTEELANDT, S., BEKAERT, M. and CLAESKENS, G. (2012). On model selection and model misspecification in causal inference. *Stat. Methods Med. Res.* **21** 7–30. MR2867536 https://doi.org/10.1177/0962280210387717

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

VON PLATO, J. (1989). De Finetti's earliest works on the foundations of probability. *Erkenntnis* **31** 263–282.

WESTREICH, D., COLE, S. R., YOUNG, J. G., PALELLA, F., TIEN, P. C., KINGSLEY, L., GANGE, S. J. and HERNÁN, M. A. (2012). The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Stat. Med.* **31** 2000–2009. MR2956032 https://doi.org/10.1002/sim.5316