Bayesian Inference for Sequential Treatments under Latent Sequential Ignorability

Alessandra Mattei[1], Federico Ricciardi[2] and Fabrizia Mealli[1]
[1]Department of Statistics, Computer Science, Applications, University of Florence, Italy
[2]Department of Statistical Science, University College London, UK

June 19, 2017

arXiv:1608.07180v2 [stat.ME] 16 Jun 2017

Abstract

We focus on causal inference for longitudinal treatments, where units are assigned to treatments at multiple time points, aiming to assess the effect of different treatment sequences on an outcome observed at a final point. A common assumption in similar studies is Sequential Ignorability (SI): treatment assignment at each time point is assumed independent of unobserved past and future potential outcomes given past observed outcomes and covariates. SI is questionable when treatment participation depends on individual choices, and treatment assignment may depend on unobservable quantities associated with future outcomes. We rely on Principal Stratification to formulate a relaxed version of SI: Latent Sequential Ignorability (LSI) assumes that treatment assignment is conditionally independent on future potential outcomes given past treatments, covariates and principal stratum membership, a latent variable defined by the joint value of observed and missing intermediate outcomes. We evaluate SI and LSI, using theoretical arguments and simulation studies to investigate the performance of the two assumptions when one holds and inference is conducted under both. Simulations show that when SI does not hold, inference performed under SI leads to misleading conclusions. Conversely, LSI generally leads to correct posterior distributions, irrespective of which assumption holds.

Keywords: Longitudinal treatments, Principal stratification, Sequential ignorablity, Rubin Causal Model.

1 Introduction

Many observational studies in different fields, including economics, social science and epidemiology, are often interested in the evaluation of causal effects of time-varying treatments, which are assigned to units sequentially over time (e.g., Robins 1986, 1989, 1997, Robins et al. 2000, Gill & Robins 2001, Lechner 2009, Achy-Brou et al. 2010, Zajonc 2012, Imai & Ratkovic 2014).

In the presence of time-varying treatments, causal inference is challenging because intermediate variables are simultaneously post-treatment outcomes and pretreatment confounders. Therefore the analysis of time-varying treatments requires methodological tools that can properly account for a growing number of intermediate variables, some of which are only partially observed, and sequential selection. In this paper we propose to face these challenges when assessing the effect of different sequences of a time-varying treatment on some final outcome observed at the end of the study.

We will frame our discussion in the context of the potential outcomes approach to causal inference, also referred to as the Rubin Causal Model (RCM, e.g., Rubin 1974, 1977, 1978, Holland 1986). A critical part of the RCM is the formulation of a treatment assignment mechanism, and this task is even more crucial in longitudinal studies. An assumption usually invoked in evaluation studies with longitudinal treatment is *Sequential Ignorability* (SI, Robins 1986), which amount to assuming that the observed treatment at a given time point is independent of future potential outcomes given past observed outcomes, past treatments and covariates up to that point. Sequential ignorability may be a reasonable assumption in various settings. For instance, in medicine, physicians may propose therapies randomly conditional on observed patient's characteristic, prognostic factors and prior treatments up to that point. In labor economics caseworkers may randomly offer training programs to participants conditional on previous training program participation up to that point and observed performances.

On the other hand, and especially for observational studies or in settings where participation in the treatment depends on individual choices, treatment assignment may depend on unobservable quantities associated with future potential outcomes as well as on unobserved past potential outcomes, even conditional on the observed history, so that the sequential ignorability assumption fails to hold. For instance, in program evaluation, subjects may decide to participate in a program at a given time point using both information on their performances under the treatments previously received (the observed outcomes), which also the experimenter can observe, as well as information on their performances under alternative unobserved treatment sequences (the missing outcomes), which may be known to subjects (maybe with some approximation) but unknown to the experimenter. In medicine, the treatment a patient decides to take at a given time point may depend on both the observed patient's history (including previous treatments and observed outcomes) as well as on some unobserved patient's characteristic related to the missing outcomes.

In order to relax SI, we rely on Principal Stratification (PS, Frangakis & Rubin 2002) and we formulate a milder version of SI that we call *Latent Sequential Ignorability* (LSI). LSI assumes that treatment assignment is conditionally independent on future potential outcomes given pre-treatment variables, past treatments, and principal strata, defined by the joint value of observed and missing intermediate outcomes up to that point. Principal strata encode personal characteristics reflected in the intermediate outcomes, therefore if intermediate outcomes are associated with future treatment and outcomes they can be viewed as a coarsened representation of the latent

unobserved structure that may affect the decision to participate in the treatment. Alternative assumptions could be considered, e.g., by gleaning from the literature on non-ignorable missing data, but we look at LSI as a valuable starting point to move forward the traditional SI assumption.

LSI has appealing features, but also raises challenging inferential issues due to the latent nature of principal strata. We propose the Bayesian approach for inference, which is particularly useful for accounting for uncertainties and for pooling information from the data in complex settings. Under SI causal estimands, such as average causal effects, are usually point identified, that is, they can be expressed as known function of the distribution of the observed data. Under LSI, some parameters may be partially or weakly identified in the sense that their posterior distributions have substantial regions of flatness (Gustafson 2010). The Bayesian approach, however, is particularly appealing to draw inference on partially or weakly identified parameters. In fact, Bayesian inference is based on the posterior distribution of the parameters of interest, which are derived by updating a prior distribution to a posterior distribution via a likelihood, irrespective of whether the parameters are fully or partially identified, and if the prior is proper, the posterior distribution will be proper, too. Bayesian analysis conducted under LSI naturally provides a framework for sensitivity analysis with respect to violations of SI, where sensitivity parameters are meaningful quantities, with a direct interpretation.

In this work we discuss and compare sequential ignorability and latent sequential ignorability, using both theoretical arguments and simulation studies in which we investigate the relative performance of the two alternative assumptions when, in turn, one holds and inference is conducted under both assumptions. We also illustrate our framework using real data on financial aids to firms to investigate the effectiveness of interests free loans on firms' employment policies. In this study firms may have access to public loans multiple times over subsequent years and our focus is on contrasting firms' performances measured in terms of employment levels at the end of the study under different treatment sequences (Pirani et al. 2013).

Throughout the article we focus on assessing causal effects of a specified longitudinal treatment on an outcome that would have been observed at the end of the study. A valuable topic for future research is the extension of our framework to the evaluation of dynamic treatment regimes, which usually describe adaptive policies that propose actions in each treatment period depending on past observations and decisions (e.g., Heckman & Navarro 2007, Hong & Raudenbush 2008, Murphy 2003, Robins 2004, Zajonc 2012).

The article is organized as follows. In Section 2 we introduce the framework and the causal estimands we focus on. In Section 3 we formally define the assignment mechanism and the critical assumptions, SI and LSI. In Section 4 we compare SI and LSI, highlighting their implications and showing how latent sequential ignorability provides a natural framework for assessing the robustness of the estimates to specific violations of the sequential ignorability assumption. In Section 5 we discuss the inferential challenges arising with longitudinal treatments, briefly reviewing the existing approaches to address them, which are mainly based on SI. We then describe the Bayesian framework for inference, a natural and appealing approach that also allows us to make comparisons between SI and LI on the same ground. In Section 6 we investigate the role and implications of the two alternative assumptions using some simulated experiments. In Section 7 we conduct causal inference under SI and LSI in the context of the illustrative case study. Finally, we conclude with a discussion in Section 8.

2 Basic Setup

In this article we will focus on a simple setup with a two-period structure and binary treatments. This simplified setting allows us to clearly describe all the conceptual issues surrounding sequential treatments, avoiding technical complications that may mask our primary objective, that is, highlighting the implications of SI and LSI and comparing inferences under the two assumptions. Indeed, the extension to more time points makes notation more complicated, but does not represent an issue for the theoretical framework, although it may raise inferential and computational challenges.

2.1 Notation

Consider a group of units indexed by $i \in \{1,\ldots,n\}$. In each of two periods, indexed by $t = 1,2$, units can be potentially assigned either an active treatment ($w_t = 1$) or a control treatment, which may be no treatment at all ($w_t = 0$). Let W_{it} denote the treatment unit i actually receives at time t: $W_{it} = 1$ if the unit is exposed to the active treatment, $W_{it} = 0$ if the unit is exposed to the control treatment. Let $\boldsymbol{W}_i = (W_{i1}, W_{i2})$. Then $\boldsymbol{W}_i \in \{(0,0),(1,0),(0,1),(1,1)\}$, that is, units can experience treatment in neither period, $\boldsymbol{W}_i = (0,0)$; only in the first period, $\boldsymbol{W}_i = (1,0)$; only in the second period, $\boldsymbol{W}_i = (0,1)$; or in both periods, $\boldsymbol{W}_i = (1,1)$. Let $\boldsymbol{W}_t$ denote the n-dimensional vector with i-th element W_{it}, which is a random vector prior to the assignment at time t, and let $\boldsymbol{w}_t$ be a realization of the random vector $\boldsymbol{W}_t$.

Let Y_{i2} denote the final outcome, which is the object of primary interest and it is measured after assignment of the final treatment, $\boldsymbol{W}_2$. After assignment to the first treatment, but prior to the assignment to the second treatment, an intermediate outcome, Y_{i1}, can be measured for each unit i. The intermediate variable we consider is the lagged outcome (or a transformation of the lagged outcome), which is a measure of the same substantive quantity as the final outcome, but measured at a previous time-point between the receipt of the first treatment and the receipt of the final treatment. This choice is compelling, since it is reasonable to believe that the lagged intermediate outcome is related to both the treatment assignment at time $t = 2$, $\boldsymbol{W}_2$, and the final outcome, Y_{i2}.

For each unit i, let $Y_{i1}(\boldsymbol{w}_1)$ denote the potential outcomes for the intermediate variable at time $t = 1$ given treatment assignment $\boldsymbol{w}_1$ in the first period, and let $Y_{i2}(\boldsymbol{w}_1, \boldsymbol{w}_2)$ denote the potential outcome for the final outcome given the entire treatment assignment sequence, $(\boldsymbol{w}_1, \boldsymbol{w}_2)$.

We make the Stable Unit Treatment Value Assumption (SUTVA, Rubin 1980), stating that potential outcomes for any unit are unaffected by the treatment assignments of other units (no interference), and that for each unit there are no different versions of treatment. Formally,

Assumption 1 *SUTVA*

If $w_{i1} = w'_{i1}$, *then* $Y_{i1}(\boldsymbol{w}_1) = Y_{i1}(\boldsymbol{w}'_1)$;
If $(w_{i1}, w_{i2}) = (w'_{i1}, w'_{i2})$, *then* $Y_{i2}(\boldsymbol{w}_1, \boldsymbol{w}_2) = Y_{i2}(\boldsymbol{w}'_1, \boldsymbol{w}'_2)$.

SUTVA allows us to write $Y_{i1}(\boldsymbol{w}_1) = Y_{i1}(w_{i1})$ and $Y_{i2}(\boldsymbol{w}_1, \boldsymbol{w}_2) = Y_{i2}(w_{i1}, w_{i2})$, therefore for each unit i there are two potential outcomes for the post-treatment intermediate variable measured after assignment to the first treatment, $Y_{i1}(0)$ and $Y_{i1}(1)$, and four potential outcomes for the final outcome, $Y_{i2}(0,0)$, $Y_{i2}(1,0)$, $Y_{i2}(0,1)$ and $Y_{i2}(1,1)$.

2.2 Causal Estimands

Causal effects on the final outcome, Y_{i2}, are defined at the unit-level as comparisons of potential outcomes for the final outcome under alternative treatment sequences. For instance, a causal effect of the treatment sequence (w_1, w_2) versus the treatment sequence (w'_1, w'_2) for a unit i is defined as a comparison between the potential outcomes $Y_{i2}(w_1, w_2)$ and $Y_{i2}(w'_1, w'_2)$. Estimands of interest may be simple differences $Y_{i2}(w_1, w_2) - Y_{i2}(w'_1, w'_2)$, but in general comparisons can take different forms. Causal effects can also be defined for collections of units. More generally, causal effects are comparisons between potential outcomes for a common set of units (Frangakis & Rubin 2002, Rubin 2005). In this article we consider the n units as a random sample from a large superpopulation, and we focus on population Average Treatment Effects (ATEs) on the final outcome, that is, the expected value of the difference between potential outcomes at time $t = 2$ under different treatment sequences. In the presence of two-period binary treatments, we have:

$$ATE_{w_1 w_2 . w'_1 w'_2} = E[Y_{i2}(w_1, w_2) - Y_{i2}(w'_1, w'_2)], \qquad \text{for } (w_1, w_2) \neq (w'_1, w'_2) \in \{0,1\}^2. \tag{1}$$

We focus on six causal effects by comparing the following treatment sequences: $(1,1)$ versus $(1,0), (0,1)$ and $(0,0)$; $(1,0)$ versus $(0,1)$ and $(0,0)$; and $(0,1)$ versus $(0,0)$.

3 The Assignment Mechanism

The fundamental problem of causal inference (Holland 1986, Rubin 1978) is that for each unit we can only observe at most one of the potential outcomes for each post-treatment variable. In our setting with two-period binary treatments, for each unit i we observe one out of two intermediate potential outcomes at time $t = 1$, i.e., $Y_{i1}^{obs} = Y_{i1}(W_{i1})$; and one out of four potential outcomes at time $t = 2$, i.e. $Y_{i2}^{obs} = Y_{i1}(W_{i1}, W_{i2})$. Potential outcomes under unassigned treatment sequences are missing: $Y_{i1}^{mis} = Y_{i1}(1 - W_{i1})$ and $\boldsymbol{Y}_{i2}^{mis} = \{Y_{i2}(1 - W_{i1}, W_{i2}), Y_{i2}(W_{i1}, 1 - W_{i2}); Y_{i2}(1 - W_{i1}, 1 - W_{i2})\}$. Therefore, inference on causal effects require to solve a missing data problem, which is particularly challenging in the presence of longitudinal treatments, even in the case with two-period binary treatments.

In order to learn about the causal effects of interest it is crucial to posit a treatment assignment mechanism. The assignment mechanism is a row-exchangeable function of all covariates and of all potential outcomes, giving the probability of any vector of treatment sequences. For each unit i, let $\boldsymbol{X}_i$ denote an observed vector of pre-treatment variables, variables that are not affected by treatments assignment. The assignment mechanism for a two-period treatment can be formally defined as follows:

$$Pr(\boldsymbol{W} \mid \boldsymbol{X}, \boldsymbol{Y}_1(0), \boldsymbol{Y}_1(1), \boldsymbol{Y}_2(0,0), \boldsymbol{Y}_2(1,0), \boldsymbol{Y}_2(0,1), \boldsymbol{Y}_2(1,1))$$

where $\boldsymbol{W}$ is a $n \times 2$ matrix with i-th row equal to (W_{i1}, W_{i2}), $\boldsymbol{X}$ is a matrix with n rows and i-th row equal to $\boldsymbol{X}_i$, and $\boldsymbol{Y}_1(w_1)$ and $\boldsymbol{Y}_2(w_1, w_2)$ are $n-$dimensional vectors with ith elements equal to $Y_{i1}(w_1)$ and $Y_{i2}(w_1, w_2)$, respectively, for $w_1 \in \{0,1\}$ and $w_2 \in \{0,1\}$.

In longitudinal settings the assignment mechanism is very complex. We consider two basic restrictions on the

assignment mechanism, assuming that it is individualistic and probabilistic. Let

$$p_i(\boldsymbol{w} \mid \boldsymbol{X}, \boldsymbol{Y}_1(0), \boldsymbol{Y}_1(1), \boldsymbol{Y}_2(0,0), \boldsymbol{Y}_2(1,0), \boldsymbol{Y}_2(0,1), \boldsymbol{Y}_2(1,1)) = \sum_{\boldsymbol{W}:\boldsymbol{W}_i=\boldsymbol{w}} Pr(\boldsymbol{W} \mid \boldsymbol{X}, \boldsymbol{Y}_1(0), \boldsymbol{Y}_1(1), \boldsymbol{Y}_2(0,0), \boldsymbol{Y}_2(1,0), \boldsymbol{Y}_2(0,1), \boldsymbol{Y}_2(1,1))$$

denote the unit-level assignment probabilities for $\boldsymbol{w} \in \{0,1\}^2$. An assignment mechanism is individualistic if

$$p_i(\boldsymbol{w} \mid \boldsymbol{X}, \boldsymbol{Y}_1(0), \boldsymbol{Y}_1(1), \boldsymbol{Y}_2(0,0), \boldsymbol{Y}_2(1,0), \boldsymbol{Y}_2(0,1), \boldsymbol{Y}_2(1,1)) = Pr(\boldsymbol{W}_i = \boldsymbol{w} \mid \boldsymbol{X}_i, Y_{i1}(0), Y_{i1}(1), Y_{i2}(0,0), Y_{i2}(1,0), Y_{i2}(0,1), Y_{i2}(1,1))$$

for all $i = 1, \ldots, n$ and $\boldsymbol{w} \in \{0,1\}^2$, and

$$Pr(\boldsymbol{W} \mid \boldsymbol{X}, \boldsymbol{Y}_1(0), \boldsymbol{Y}_1(1), \boldsymbol{Y}_2(0,0), \boldsymbol{Y}_2(1,0), \boldsymbol{Y}_2(0,1), \boldsymbol{Y}_2(1,1)) \propto \prod_i \prod_{\boldsymbol{w} \in \{0,1\}^2} Pr(\boldsymbol{W}_i = \boldsymbol{w} \mid \boldsymbol{X}_i, Y_{i1}(0), Y_{i1}(1), Y_{i2}(0,0), Y_{i2}(1,0), Y_{i2}(0,1), Y_{i2}(1,1))^{\mathbf{1}\{\boldsymbol{W}_i = \boldsymbol{w}\}}$$

for $(\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}_1(0), \boldsymbol{Y}_1(1), \boldsymbol{Y}_2(0,0), \boldsymbol{Y}_2(1,0), \boldsymbol{Y}_2(0,1), \boldsymbol{Y}_2(1,1)) \in \mathbb{A}$, for some set $\mathbb{A}$, and zero otherwise.

An assignment mechanism is probabilistic if

$$0 < p_i(\boldsymbol{w} \mid \boldsymbol{X}_i, Y_{i1}(0), Y_{i1}(1), Y_{i2}(0,0), Y_{i2}(1,0), Y_{i2}(0,1), Y_{i2}(1,1)) < 1,$$

for all $i = 1, \ldots, n$, and $\boldsymbol{w} \in \{0,1\}^2$.

Even under these restrictions, the assignment mechanism still remains complex, because it depends on a large number of missing values, Y_{i1}^{mis} and $\boldsymbol{Y}_{i2}^{mis}$, for all i. In order to reduce the complexity of the assignment *mechanism*, we now formulate some assumptions, which allow us to characterize longitudinal observational studies and draw inference on the causal estimands of interest. To this end, it is useful to factorize the unit-level assignment probabilities as product of the assignment probabilities at time $t = 1$ and the conditional assignment probabilities at time $t = 2$ given the treatment received at time one. Formally, by the law of total probability, we have

$$\begin{aligned} Pr(\boldsymbol{W}_i \mid \boldsymbol{X}_i, Y_{i1}(0), Y_{i1}(1), Y_{i2}(0,0), Y_{i2}(1,0), Y_{i2}(0,1), Y_{i2}(1,1)) = \\ Pr(W_{i1} \mid \boldsymbol{X}_i, Y_{i1}(0), Y_{i1}(1), Y_{i2}(0,0), Y_{i2}(1,0), Y_{i2}(0,1), Y_{i2}(1,1)) \times \\ Pr(W_{i2} \mid W_{i1}, \boldsymbol{X}_i, Y_{i1}(0), Y_{i1}(1), Y_{i2}(0,0), Y_{i2}(1,0), Y_{i2}(0,1), Y_{i2}(1,1)). \end{aligned}$$

Much of the literature on time-varying treatments copes with the complications arising in the presence of sequential treatments by assuming that the assignment mechanism is sequentially ignorable (Robins 1986):

Assumption 2 *Sequential Ignorability (SI)*

$$Pr(W_{i1} \mid \boldsymbol{X}_i, Y_{i1}(0), Y_{i1}(1), Y_{i2}(0,0), Y_{i2}(1,0), Y_{i2}(0,1), Y_{i2}(1,1)) = Pr(W_{i1} \mid \boldsymbol{X}_i) \quad (2)$$

$$Pr\left(W_{i2} \mid \boldsymbol{X}_i, W_{i1}, Y_{i1}^{obs}, Y_{i2}(0,0), Y_{i2}(1,0), Y_{i2}(0,1), Y_{i2}(1,1)\right) = Pr\left(W_{i2} \mid \boldsymbol{X}_i, W_{i1}, Y_{i1}^{obs}\right) \quad (3)$$

SI implies that treatment assignment at each time point is independent of all future potential outcomes given past observed outcomes, treatments and covariates.

SI guarantees that, within cells defined by the pre-treatment covariates, the mean of the potential outcomes under a specific treatment sequence can be estimated from the observed data as weighted average of the means of the observed final outcome under that treatment sequence across groups defined by the observed intermediate outcome, with weights that depend on the distribution of the observed intermediate outcome. Formally, under SI

$$E\left[Y_{i2}(w_1,w_2) \mid \boldsymbol{X}_i\right] = \int E\left[Y_{i2}^{obs} \mid W_{i1}=w_1, W_{i2}=w_2, Y_{i1}^{obs}=y_1, \boldsymbol{X}_i\right] d\, F_{Y_{i1}^{obs}|W_{i1}=w_1,\boldsymbol{X}_i}(y_1),$$

where $F_{Y_{i1}^{obs}|W_{i1}=w_1,\boldsymbol{X}_i}(\cdot)$ is the conditional cumulative distribution function of the intermediate outcome, Y_1^{obs}, given the observed treatment at time $t=1$ and pre-treatment covariates.

It is worth noting that SI defines the assignment mechanism at each time point separately and independently of the other time points. Essentially the underlying idea is that at each time point a new study has been conducted, for which an assignment mechanism must be posited, and SI implies that at every time the treatment is as if randomized with probabilities depending on the observed history. Although SI allows one to easily identify and estimate the conditional expectation of the potential outcomes of interest, it does not permit to reconstruct the assignment mechanism underlying the longitudinal study in its entirety, that is, the joint conditional probability of $\boldsymbol{W}_i$ given all the potential outcomes and covariates. To this end we can introduce a different ignorability assumption, which is highly related to SI:

Assumption 3 *Sequential Ignorability of Longitudinal Treatment Assignment (SIL)*

$$\begin{aligned} &Pr(\boldsymbol{W}_i \mid \boldsymbol{X}_i, Y_{i1}(0), Y_{i1}(1), Y_{i2}(0,0), Y_{i2}(1,0), Y_{i2}(0,1), Y_{i2}(1,1)) = \\ &Pr(W_{i1} \mid \boldsymbol{X}_i) \times Pr\left(W_{i2} \mid W_{i1}, Y_{i1}^{obs}, \boldsymbol{X}_i\right). \end{aligned}$$

Assumption 3 amounts to assuming that treatment assignment at each time point is independent of past missing potential outcomes and all future potential outcomes given past observed outcomes, treatments and covariates. Assumption 3 is slightly stronger than Assumption 2, because it implies Assumption 2 but the converse is not true: Assumption 2 ignores the relationship between treatment assignment at time $t=2$ and past missing potential outcomes, only requiring that the assignment mechanism at time $t=2$ is independent of all future potential outcomes conditional on the observed history. Nevertheless Assumptions 2 and 3 have the same implications from an inferential perspective. For this reason, although Assumption 2 is weaker than Assumption 3, in practice it is difficult that a convincing argument can be made for the weaker Assumption 2 without the argument being equally cogent for the stronger Assumption 3.

Sequential ignorability assumptions may be reasonable in various settings, including longitudinal observational studies where it is reasonable to believe that treatments are sequentially assigned using only the observed information (e.g., Zajonc 2012). However, as in single point observational studies, where the usually made strong ignorability assumption may fail to hold due to the presence of unobserved confounders associated with both the potential outcomes and the treatment indicator (Rosenbaum & Rubin 1983, Rosenbaum 1987, Imbens 2003,

Ichino et al. 2008), here sequential ignorability may be arguable due to the presence of time-varying unobserved confounder factors. The key insight is that the joint potential values of the intermediate outcome at time $t = 1$, $(Y_{i1}(0), Y_{i1}(1))$, may represent an accurate summary of the unobserved variables related to both treatment assignment at time $t = 2$ and the final outcome, due to which sequential ignorability assumptions do not hold.

Motivated by this intuition, we use the concept of principal stratification (Frangakis & Rubin 2002) to define a new assumption on the longitudinal assignment mechanism, which may be a valuable alternative to sequential ignorability assumptions when they are assumed to fail in some specific and meaningful ways. The joint potential values of the intermediate outcome at time $t = 1$, $(Y_{i1}(0), Y_{i1}(1))$, defines a classification of units into principal strata. Principal stratification *per se* does not require that the intermediate outcome is binary or categorical. Recent work has indeed considered the application of principal stratification in the presence of continuous post-treatment variables (e.g., Schwartz et al. 2011). Nevertheless, continuous intermediate variables introduce serious challenges to principal stratification analysis. Specifically continuous intermediate outcomes induce an infinite number of possible principal strata, leading to substantial complications in both inference and interpretation. In order to avoid additional complications, which may mask our primary objectives, here we consider a binary intermediate variable. Thus, the (basic) principal stratification with respect to the binary intermediate outcome Y_1 classifies units into four groups according to the joint potential values of Y_1, $Y_{i1}(0)$ and $Y_{i1}(1)$: $00 = \{i : (Y_{i1}(0) = 0, Y_{i1}(1) = 0)\}$; $01 = \{i : (Y_{i1}(0) = 0, Y_{i1}(1) = 1)\}$; $10 = \{i : (Y_{i1}(0) = 1, Y_{i1}(1) = 0)\}$; and $11 = \{i : (Y_{i1}(0) = 1, Y_{i1}(1) = 1)\}$. Let G_i denote the principal stratum membership for unit i, with $i = (1, \ldots, n)$, then $G_i \equiv (Y_{i1}(0), Y_{i1}(1)) \in \{00, 01, 10, 11\}$.

For instance, in our illustrative example, the intermediate outcome is an indicator variable taking on value one if a firm hires new staff between the assignment to the first treatment and the assignment to the second treatment. Therefore, for example, principal stratum 11 includes firms that would hire new staff irrespective of their treatment assignment at time $t = 1$ (see Section 7 for further details).

Principal stratum membership G_i is not affected by treatment assignment at time $t = 1$, W_{i1}, so it only reflects characteristics of unit i. Therefore, principal strata can be viewed as a representation of the latent unobserved structure that may influence the decision to participate in the treatment at a future time point.

Based on principal stratification, we introduce a Latent Sequential Ignorability (LSI) assumption, where the word *latent* indicates that treatment assignment is conditionally independent on future potential outcomes conditionally on pre-treatment covariates, past treatments and the latent indicator for principal stratum membership. In fact we cannot, in general, observe the principal stratum which a unit belongs to, because principal strata are defined by the joint values of observed and missing intermediate outcomes. In other words, LSI is a form of latent ignorability (Frangakis & Rubin 1999), in that it conditions on variables that are (at least partially) unobserved or latent. Formally:

Assumption 4 *Latent Sequential Ignorability (LSI)*

$$Pr(\boldsymbol{W}_i \mid \boldsymbol{X}_i, Y_{i1}(0), Y_{i1}(1), Y_{i2}(0,0), Y_{i2}(1,0), Y_{i2}(0,1), Y_{i2}(1,1)) =$$
$$Pr(W_{i1} \mid \boldsymbol{X}_i) \times Pr(W_{i2} \mid W_{i1}, Y_{i1}(0), Y_{i1}(1), \boldsymbol{X}_i).$$

LSI is a relaxed version of SIL (Assumption 3): SIL implies LSI, therefore SIL is a stronger assumption. LSI can

be equivalently formulated as follows

$$Pr(W_{i1} \mid \boldsymbol{X}_i, Y_{i1}(0), Y_{i1}(1), Y_{i2}(0,0), Y_{i2}(1,0), Y_{i2}(0,1), Y_{i2}(1,1)) = Pr(W_{i1} \mid \boldsymbol{X}_i) \quad (4)$$

$$Pr(W_{i2} \mid \boldsymbol{X}_i, W_{i1}, Y_{i1}(0), Y_{i1}(1), Y_{i2}(0,0), Y_{i2}(1,0), Y_{i2}(0,1), Y_{i2}(1,1)) = Pr(W_{i2} \mid \boldsymbol{X}_i, W_{i1}, Y_{i1}(0), Y_{i1}(1)) \quad (5)$$

This formulation of LSI makes it clear the critical difference between SI (Assumption 2) and LSI. Although SI and LSI both assume that the assignment mechanism at time $t = 1$ is ignorable given the set of observable variables, $\boldsymbol{X}_i$ (see Equation (2) and Equation (4)), SI and LSI impose different restrictions on the assignment mechanism at time $t = 2$: standard sequential ignorability implies that it is ignorable given the observable past history, whereas LSI requires that it is ignorable given the observable past history and the missing intermediate outcomes.

LSI implies that

$$E\left[Y_{i2}(w_1, w_2) \mid \boldsymbol{X}_i\right] = \int E\left[Y_{i2}^{obs} \mid W_{i1} = w_1, W_{i2} = w_2, G_i = g, \boldsymbol{X}_i\right] dF_{G_i|\boldsymbol{X}_i}(g),$$

where $F_{G_i|\boldsymbol{X}_i}(\cdot)$ is the conditional cumulative distribution function of the principal stratum membership, G, given pre-treatment covariates. Therefore if principal stratum membership were observed, under LSI within cell defined by the covariates, $E\left[Y_{i2}(w_1, w_2) \mid \boldsymbol{X}_i\right]$ could be derived as the weighted average of the means of the observed outcome for units with $W_{i1} = w_1$ and $W_{i2} = w_2$ across principal strata with weights that depends on the conditional distribution of principal strata given covariates. In practice, principal stratum membership is generally unobserved, therefore inference under LSI raises non trivial challenges (see Section 5.1 for details on inference under LSI).

4 Assessing Sequential Ignorability through Latent Sequential Ignorability

In this section we investigate the role of LSI (Assumption 4) in causal inference for sequential treatment. Let first consider the relationship between SIL (Assumption 3) and LSI (Assumption 4). LSI is a relaxed version of SIL and for this reason SIL can be viewed as a special case of LSI. Therefore, in order to compare SIL with LSI and to investigate which one is more appropriate for a given problem at hand, we rely on the relationship between SIL and LSI when SIL holds.

Under SIL, treatment assignment at $t = 2$ does not depend on the missing intermediate potential outcomes, implying that treatment assignment probabilities are homogeneous across some principal strata, conditionally on the treatment assigned at $t = 1$ and covariates. Specifically, SIL implies that the assignment probabilities of W_{i2} in principal strata sharing the same value for the observed intermediate outcome that is, the intermediate outcome under the treatment assigned at time 1, are the same. Formally, under SIL, for each $w_1 = 0, 1$ and $y_1 = 0, 1$, we have

$$Pr(W_{i2}|\boldsymbol{X}_i, W_{i1} = w_1, Y_{i1}^{obs} = y_1, Y_{i1}^{mis}) = Pr(W_{i2}|\boldsymbol{X}_i, W_{i1} = w_1, Y_{i1}^{obs} = y_1).$$

Therefore, if SIL holds we have:

$$\begin{aligned} Pr(W_{i2}|\boldsymbol{X}_i, W_{i1}=0, G_i=00) &= Pr(W_{i2}|\boldsymbol{X}_i, W_{i1}=0, G_i=01), \\ Pr(W_{i2}|\boldsymbol{X}_i, W_{i1}=0, G_i=10) &= Pr(W_{i2}|\boldsymbol{X}_i, W_{i1}=0, G_i=11), \\ Pr(W_{i2}|\boldsymbol{X}_i, W_{i1}=1, G_i=00) &= Pr(W_{i2}|\boldsymbol{X}_i, W_{i1}=1, G_i=10), \\ Pr(W_{i2}|\boldsymbol{X}_i, W_{i1}=1, G_i=01) &= Pr(W_{i2}|\boldsymbol{X}_i, W_{i1}=1, G_i=11). \end{aligned} \tag{6}$$

Under SI (Assumption 2) the assignment probabilities of W_{i2} only depend on the observed intermediate outcomes conditionally on the treatment assigned at $t=1$ and covariates: $Pr(W_{i2}|\boldsymbol{X}_i, W_{i1}=w_1, Y_{i1}^{obs})$, therefore they can be ignored in drawing inference on the causal effects of interest. If SI does not hold, but LSI holds, ignoring the assignment probabilities of W_{i2}, $Pr(W_{i2}|\boldsymbol{X}_i, W_{i1}=w_1, G_i=g)$, does not, in general, lead to a valid analysis. This result suggests that we can investigate the robustness of the estimated causal effects with respect to violations of the sequential ignorability assumptions, using the assignment probabilities under LSI, $Pr(W_{i2}|\boldsymbol{X}_i, W_{i1}=w_1, G_i=g)$, as sensitivity parameters.

If principal strata encode characteristics of the units that are associated with the treatment assigned at time $t=2$ and possibly with the final outcome, i.e., LSI holds but neither SIL nor SI holds, inference under LSI is expected to show evidence against at least one of the equalities in Equation (6), and SI/SIL and LSI are expected to lead to substantially different inferential conclusions on the causal effects of interest. Conversely, if we find that treatment assignment probabilities are homogeneous across principal strata according to the equalities in Equation (6), then causal inference under sequential ignorability is more defensible.

In this sense, LSI naturally provides a framework for sensitivity analysis with respect to violations of sequential ignorability: looking at the inferential results on the assignment probabilities under LSI we can get some insight on the plausibility of the sequential ignorability assumptions. This framework for sensitivity analysis is in line with the existing approaches in the literature to sensitivity analysis with respect to violations of the unconfoundness assumption, usually made in single time observational studies (Rosenbaum & Rubin 1983, Rosenbaum 1987, Imbens 2003, Ichino et al. 2008, Ding & VanderWeele 2016), where the robustness of the estimated causal effects with respect to the unconfoundness assumption is generally assessed focusing on its violations due to the presence of unobserved covariates that are correlated both with the potential outcomes and with the treatment indicator. In those settings, sensitivity parameters are quantities characterizing the distribution of the unobserved covariates and their association with the potential outcomes and with the treatment indicator, but they do not generally have a substantial meaning. In our framework, sensitivity parameters are meaningful quantities with a direct interpretation: they are the assignment probabilities for specific sub-population of units.

5 Inference

Under SI and SIL average causal effects are point identified, i.e, they can be expressed as known function of the distribution of the observed data, since different effect values cannot correspond to the same distribution of the observables. Therefore, ideally, we could estimate average treatment effects non-parametrically. In practice, data are often sparse and high dimensional, and model assumptions are usually introduced. Methods usually applied

Table 1: Group classification based on observed data $O(W_{i1}, Y_{i1}^{obs})$, associated data pattern and latent principal strata.

Observed group			Latent group	
$O(W_{i1}, Y_{i1}^{obs})$	W_{i1}	Y_{i1}^{obs}	G_i	
$O(0,0)$	0	0	00	01
$O(0,1)$	0	1	10	11
$O(1,0)$	1	0	00	10
$O(1,1)$	1	1	01	11

to estimate causal effects of longitudinal treatment under SI (Assumption 2) include the G-computation algorithm formula (Robins 1986), inverse probability of treatment weighting estimation of marginal structural models (Robins 1989), and G-estimation of structural nested models (Robins 1999). The three methods would give identical estimates of the treatment effects if a non-parametric approach to inference or saturated marginal structural models/structured nested models were used, but under model assumptions they generally provide different estimates, depending on the specific parametric assumptions that are introduced. The G-formula requires to specify many models, often raising model-compatibility issues. Marginal structural models (MSMs) and structured nested models, which have received increasing attention in the last years, require to specify models for marginal potential outcomes ($Y_2(w_1, w_2)$ for each $(w_1, w_2) \in \{0,1\}^2$ in our setting) and for the causal effects, which may assume, e.g., constant treatment effects, additivity and so on. Moreover inferential methods based on inverse probability of treatment weighting require to also specify a model for the probability of treatment. These assumptions may be critical because model misspecification may lead to biased estimates of the treatment effects even if the identifiability conditions hold.

Under LSI the average causal effects are generally not point identified, due to the latent nature of the principal strata. In our setting, we can only observe four groups based on the treatment actually received at time $t = 1$, W_{i1}, and the observed value of the intermediate outcome, Y_{i1}^{obs}, and each of them comprises a mixture of two principal strata, as shown in the last two columns in Table 1. In the principal stratification literature, structural or modeling assumptions are typically invoked (e.g., Imbens & Rubin 1997, Mattei & Mealli 2007, Schwartz et al. 2011). Monotonicity and exclusion restriction assumptions, usually used in experimental studies with noncompliance, may be questionable in longitudinal settings. Depending on the substantive empirical setting, other structural or modeling assumptions can be introduced. In this paper we prefer to avoid structural assumptions, which may make the comparison between SI/SIL and LSI unfair or strongly depending on some specific assumption, and we opt for a model-based approach for inference.

Following the literature on principal stratification, models for potential outcomes are specified conditional on covariates and principal strata (see Section 5.1 for further details). Again, distributional assumptions may be critical. Nevertheless in our opinion this model-based approach is very flexible, and in some settings model assumptions on the conditional distributions of potential outcomes may be less demanding than model assumptions on the marginal distributions of potential outcomes and on the causal effects. In order to make the comparison between SI/SIL and LSI as fair as possible, the same model-based approach is used under SI/SIL, although we will also show results from G-methods under SI. An advantage of this model-based approach is that it allows us to directly get information on the heterogeneity of the effects with respect to principal strata both under SI/SIL and

LSI.

5.1 Bayesian Inference

We adopt a Bayesian approach to inference, which is particularly suitable for model-based causal inference. The Bayesian perspective appears to be particularly appropriate for addressing problems of causal inference because it treats the uncertainty in the missing potential outcomes in the same way that it treats the uncertainty in the unknown parameters. A Bayesian approach explicitly deals with the different sources of uncertainty, treating them separately. Also in a Bayesian framework, we can be formally clear about the role played by the treatment assignment mechanism and the complications that raise in drawing inference for sequential treatments under LSI (Rubin 1978, Imbens & Rubin 1997). From a Bayesian perspective, all inferences are based on the posterior distribution of the causal estimands, defined as functions of observed and unobserved potential outcomes, or sometimes as functions of model parameters (Rubin 1978). Because with proper prior distributions, posterior distributions are always proper, from a Bayesian perspective, there is no conceptual difference between fully and partially/weakly identified parameters. Weak identifiability is usually reflected in the flatness of the posterior distribution (Gustafson 2010). Therefore the Bayesian approach appears to be a natural and appealing inferential approach to make comparisons between SI/SIL and LSI on the same ground.

Bayesian inference considers the observed values to be realizations of random variables and the missing values to be unobserved random variables, starting from the joint probability distributions of all random variables for all units:

$$p(\boldsymbol{Y}_1(0),\boldsymbol{Y}_1(1),\boldsymbol{Y}_2(0,0),\boldsymbol{Y}_2(0,1),\boldsymbol{Y}_2(1,0),\boldsymbol{Y}_2(1,1),\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{X}).$$

We assume this distribution is unit exchangeable, that is, invariant under a permutation of the indexes, then de Finetti's theorem (de Finetti 1937, 1964) implies that there exists a vector of parameters $\boldsymbol{\theta}$, which is a random variable itself, with prior distribution $p(\boldsymbol{\theta})$, such that $\boldsymbol{Y}_1(0),\boldsymbol{Y}_1(1),\boldsymbol{Y}_2(0,0),\boldsymbol{Y}_2(0,1),\boldsymbol{Y}_2(1,0),\boldsymbol{Y}_2(1,1),\boldsymbol{W}_1,\boldsymbol{W}_2$ and $\boldsymbol{X}$ consist of independent and identically distributed random variables given $\boldsymbol{\theta}$. Thus,

$$\begin{aligned} &p(\boldsymbol{Y}_2(0,0),\boldsymbol{Y}_2(0,1),\boldsymbol{Y}_2(1,0),\boldsymbol{Y}_2(1,1),\boldsymbol{Y}_1(0),\boldsymbol{Y}_1(1),\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{X}) = \\ &\quad \int \textstyle\prod_i p(Y_{i2}(0,0),Y_{i2}(0,1),Y_{i2}(1,0),Y_{i2}(1,1),Y_{i1}(0),Y_{i1}(1),W_{i1},W_{i2},\boldsymbol{X}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}, \end{aligned} \tag{7}$$

and the posterior distribution of $\boldsymbol{\theta}$ can be written as

$$\begin{aligned} &p(\boldsymbol{\theta}|\boldsymbol{Y}_2^{obs},\boldsymbol{Y}_1^{obs},\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{X}) \propto \\ &\quad p(\boldsymbol{\theta}) \times \int\int \prod_i p(Y_{i2}(0,0),Y_{i2}(0,1),Y_{i2}(1,0),Y_{i2}(1,1),Y_{i1}(0),Y_{i1}(1),W_{i1},W_{i2},\boldsymbol{X}_i|\boldsymbol{\theta})d\boldsymbol{Y}_{i2}^{mis}dY_{i1}^{mis} = \\ &\quad p(\boldsymbol{\theta}) \times \\ &\quad \int\int \prod_i \Big[p(\boldsymbol{X}_i|\boldsymbol{\theta}) \times p(Y_{i1}(0),Y_{i1}(1)|\boldsymbol{X}_i;\boldsymbol{\theta}) \times p(Y_{i2}(0,0),Y_{i2}(0,1),Y_{i2}(1,0),Y_{i2}(1,1)|Y_{i1}(0),Y_{i1}(1),\boldsymbol{X}_i;\boldsymbol{\theta}) \times \\ &\qquad p(W_{i1}|Y_{i2}(0,0),Y_{i2}(0,1),Y_{i2}(1,0),Y_{i2}(1,1),Y_{i1}(0),Y_{i1}(1),\boldsymbol{X}_i;\boldsymbol{\theta}) \times \\ &\qquad p(W_{i2}|Y_{i2}(0,0),Y_{i2}(0,1),Y_{i2}(1,0),Y_{i2}(1,1),Y_{i1}(0),Y_{i1}(1),\boldsymbol{X}_i,W_{i1};\boldsymbol{\theta})d\boldsymbol{Y}_{i2}^{mis}dY_{i1}^{mis}\Big]. \end{aligned} \tag{8}$$

The assumptions on the assignment mechanism are crucial to draw inference on the causal estimands. Under latent sequential ignorability (Assumption 4), within cells defined by the values of pre-treatment variables $\boldsymbol{X}_i$, the treatment at time $t=1$ is assigned independently of the relevant post-treatment variables, $Y_{i1}(w_1)$ and $Y_{i2}(w_1,w_2)$, $w_1=0,1$, $w_2=0,1$, and the treatment at time $t=2$ is assigned independently of the final potential outcomes, $Y_{i2}(w_1,w_2)$, $w_1=0,1$, $w_2=0,1$, conditional on the treatment assigned at time $t=1$, W_{i1}, and the principal strata defined by $(Y_{i1}(0),Y_{i1}(1))$. Therefore, under LSI the posterior distribution of $\boldsymbol{\theta}$ becomes

$$
\begin{aligned}
&p(\boldsymbol{\theta}|\boldsymbol{Y}_2^{obs},\boldsymbol{Y}_1^{obs},\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{X}) \propto p(\boldsymbol{\theta})\times \\
&\quad\int\int\prod_i\big[p(\boldsymbol{X}_i|\boldsymbol{\theta})\times p(Y_{i1}(0),Y_{i1}(1)|\boldsymbol{X}_i;\boldsymbol{\theta})\times p(Y_{i2}(0,0),Y_{i2}(0,1),Y_{i2}(1,0),Y_{i2}(1,1)|Y_{i1}(0),Y_{i1}(1),\boldsymbol{X}_i;\boldsymbol{\theta})\times \\
&\qquad p(W_{i1}|\boldsymbol{X}_i;\boldsymbol{\theta})\times p(W_{i2}|Y_{i1}(0),Y_{i1}(1),\boldsymbol{X}_i,W_{i1};\boldsymbol{\theta})d\boldsymbol{Y}_{i2}^{mis}dY_{i1}^{mis}\big]. \qquad (9)
\end{aligned}
$$

Equation (9) further simplifies under SIL (Assumption 3), which implies that the treatment at time $t=2$ is assigned independently of both missing intermediate potential outcomes Y_{i1}^{mis} and final potential outcomes, $Y_{i2}(w_1,w_2)$, $w_1=0,1$, $w_2=0,1$, conditional on the pre-treatment variables, $\boldsymbol{X}_i$, the treatment assigned at time $t=1$, W_{i1}, and the past observed potential outcomes, Y_{i1}^{obs}:

$$
\begin{aligned}
&p(\boldsymbol{\theta}|\boldsymbol{Y}_2^{obs},\boldsymbol{Y}_1^{obs},\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{X}) \propto p(\boldsymbol{\theta})\times \\
&\quad\int\int\prod_i\big[p(\boldsymbol{X}_i|\boldsymbol{\theta})\times p(Y_{i1}(0),Y_{i1}(1)|\boldsymbol{X}_i;\boldsymbol{\theta})\times p(Y_{i2}(0,0),Y_{i2}(0,1),Y_{i2}(1,0),Y_{i2}(1,1)|Y_{i1}(0),Y_{i1}(1),\boldsymbol{X}_i;\boldsymbol{\theta})\times \\
&\qquad p(W_{i1}|\boldsymbol{X}_i;\boldsymbol{\theta})\times p(W_{i2}|Y_{i1}^{obs},\boldsymbol{X}_i,W_{i1};\boldsymbol{\theta})d\boldsymbol{Y}_{i2}^{mis}dY_{i1}^{mis}\big]. \qquad (10)
\end{aligned}
$$

The right hand of Equation (10) is also the posterior distribution of $\boldsymbol{\theta}$ under SI (Assumption 2). It is worth noting that, under the assumption that the parameters governing the distributions under the integral sign in Equations (9) and (10) are a priori distinct and independent from each other (Rubin 1978), we can ignore the distributions $p(\boldsymbol{X}_i|\boldsymbol{\theta}_X)$ and $p(W_{i1}|\boldsymbol{X}_i;\boldsymbol{\theta})$ in drawing Bayesian inference on the relevant estimateds. If SI holds, Bayesian causal inference does not even require to model the distribution of the treatment at time $t=2$, $p(W_{i2}|Y_{i1}^{obs}\boldsymbol{X}_i,W_{i1};\boldsymbol{\theta})$ (Rubin 1978, Zajonc 2012), although we decided to model it in the analyses below to better describe and discuss the role of LSI and SI/SIL in longitudinal studies.

Throughout the article we assume that conditional on $\boldsymbol{X}_i$ and $\boldsymbol{\theta}$, the four outcomes $Y_{i2}(0,0)$, $Y_{i2}(0,1)$, $Y_{i2}(1,0)$, $Y_{i2}(1,1)$ are independent. Data are not informative about the partial association structure between final potential outcomes, because $Y_{i2}(0,0)$, $Y_{i2}(0,1)$, $Y_{i2}(1,0)$, $Y_{i2}(1,1)$ are never jointly observed, but the independence assumption has little inferential effect if we regard the n units in the study as a random sample from a super-population and we focus on super-population causal estimands that do not depend on the association structure between the final potential outcomes. Indeed, the causal estimands of primary interest here, the average causal effects in Equation (1) are super-population causal estimands, which are free of the association structure between the final potential outcomes (Imbens & Rubin 1997, 2015, Chapter 6, pp. 98-101).

Let $O(W_{i1},Y_{i1}^{obs},W_{i2})$ denote the observed group defined by the observed variables W_{i1}, Y_{i1}^{obs}, and W_{i2}, and recall that $G_i\equiv(Y_{i1}(0),Y_{i1}(1))\in\{00,01,10,11\}$. Let $\pi_{ig}=p(G_i=g|\boldsymbol{X}_i;\boldsymbol{\theta})$, $h_{ig}^{w_1}=p(W_{i2}|G_i=g,\boldsymbol{X}_i,W_{i1}=w_1;\boldsymbol{\theta})$ and $f_{ig}^{w_1,w_2}=p(Y_{i2}(w_1,w_2)|G_i=g,\boldsymbol{X}_i;\boldsymbol{\theta})$, $g=00,01,10,11$, $w_1=0,1$, $w_2=0,1$. Then performing the integration in

Equation (9), under LSI the posterior distribution of $\boldsymbol{\theta}$ given the observed data can be written as follows:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{Y}_2^{obs},\boldsymbol{Y}_1^{obs},\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{X}) \propto\ & p(\boldsymbol{\theta})\times \\
& \prod_{i\in O(0,0,0)}\left[\pi_{i00}(1-h_{i00}^{0})f_{i00}^{0,0}+\pi_{i01}(1-h_{i01}^{0})f_{i01}^{0,0}\right]\times\prod_{i\in O(0,0,1)}\left[\pi_{i00}h_{i00}^{0}f_{i00}^{0,1}+\pi_{i01}h_{i01}^{0}f_{i01}^{0,1}\right]\times \\
& \prod_{i\in O(0,1,0)}\left[\pi_{i10}(1-h_{i10}^{0})f_{i10}^{0,0}+\pi_{i11}(1-h_{i11}^{0})f_{i11}^{0,0}\right]\times\prod_{i\in O(0,1,1)}\left[\pi_{i10}h_{i10}^{0}f_{i10}^{0,1}+\pi_{i11}h_{i11}^{0}f_{i11}^{0,1}\right]\times \\
& \prod_{i\in O(1,0,0)}\left[\pi_{i00}(1-h_{i00}^{1})f_{i00}^{1,0}+\pi_{i10}(1-h_{i10}^{1})f_{i10}^{1,0}\right]\times\prod_{i\in O(1,0,1)}\left[\pi_{i00}h_{i00}^{1}f_{i00}^{1,1}+\pi_{i10}h_{i10}^{1}f_{i10}^{1,1}\right]\times \\
& \prod_{i\in O(1,1,0)}\left[\pi_{i01}(1-h_{i01}^{1})f_{i01}^{1,0}+\pi_{i11}(1-h_{i11}^{1})f_{i11}^{1,0}\right]\times\prod_{i\in O(1,1,1)}\left[\pi_{i01}h_{i01}^{1}f_{i01}^{1,1}+\pi_{i11}h_{i11}^{1}f_{i11}^{1,1}\right]. \qquad (11)
\end{aligned}
$$

Therefore model-based Bayesian inference under LSI requires to specify three models: (1) the model for principal strata conditional on covariates, π_{ig};(2) the model for treatment assigned at time $t=2$ conditional on principal strata, past treatment and covariates, $h_{ig}^{w_1}$; and (3) the model for final potential outcomes conditional on principal strata and covariates, $f_{ig}^{w_1,w_2}$.

Let $h_{iy_1}^{w_1}=p(W_{i2}|Y_{i1}(w_1)=y_1,\boldsymbol{X}_i,W_{i1}=w_1;\boldsymbol{\theta})$, $y_1=0,1$, $w_1=0,1$. Performing the integration in Equation (10), under SI/SIL the posterior distribution of $\boldsymbol{\theta}$ given the observed data can be written as follows:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{Y}_2^{obs},\boldsymbol{Y}_1^{obs},\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{X}) \propto\ & p(\boldsymbol{\theta})\times \\
& \prod_{i\in O(0,0,0)}(1-h_{i0}^{0})\left[\pi_{i00}f_{i00}^{0,0}+\pi_{i01}f_{i01}^{0,0}\right]\times\prod_{i\in O(0,0,1)}h_{i0}^{0}\left[\pi_{i00}f_{i00}^{0,1}+\pi_{i01}f_{i01}^{0,1}\right]\times \\
& \prod_{i\in O(0,1,0)}(1-h_{i1}^{0})\left[\pi_{i10}f_{i10}^{0,0}+\pi_{i11}f_{i11}^{0,0}\right]\times\prod_{i\in O(0,1,1)}h_{i1}^{0}\left[\pi_{i10}f_{i10}^{0,1}+\pi_{i11}f_{i11}^{0,1}\right]\times \\
& \prod_{i\in O(1,0,0)}(1-h_{i0}^{1})\left[\pi_{i00}f_{i00}^{1,0}+\pi_{i10}f_{i10}^{1,0}\right]\times\prod_{i\in O(1,0,1)}h_{i0}^{1}\left[\pi_{i00}f_{i00}^{1,1}+\pi_{i10}f_{i10}^{1,1}\right]\times \\
& \prod_{i\in O(1,1,0)}(1-h_{i1}^{1})\left[\pi_{i01}f_{i01}^{1,0}+\pi_{i11}f_{i11}^{1,0}\right]\times\prod_{i\in O(1,1,1)}h_{i1}^{1}\left[\pi_{i01}f_{i01}^{1,1}+\pi_{i11}f_{i11}^{1,1}\right]. \qquad (12)
\end{aligned}
$$

Now, define $\pi_{iy_1}=p(Y_{i1}(w_1)=y_1|\boldsymbol{X}_i,\boldsymbol{\theta})$, $y_1=0,1$ and $f_{iy_1}^{w_1,w_2}=p(Y_{i2}(w_1,w_2)|Y_{i1}(w_1)=y_1,\boldsymbol{X}_i;\boldsymbol{\theta})$, $y_1=0,1$, $w_1=0,1$, $w_2=0,1$. Then, taking the sums in the brackets on the right hand of Equation (12), that is, marginalizing over the missing intermediate outcome, we have that $\pi_{i0y_1}f_{i01}^{w_1,w_2}+\pi_{i1y_1}f_{i1y_1}^{w_1,w_2}=\pi_{iy_1}f_{iy_1}^{w_1,w_2}$. Therefore under SI/SIL, the posterior distribution of $\boldsymbol{\theta}$ given the observed data in Equation (12) can be also written as follows:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{Y}_2^{obs},\boldsymbol{Y}_1^{obs},\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{X}) \propto\ & p(\boldsymbol{\theta})\times \\
& \prod_{i\in O(0,0,0)}(1-h_{i0}^{0})\pi_{i0}f_{i0}^{0,0}\times\prod_{i\in O(0,0,1)}h_{i0}^{0}\pi_{i0}f_{i0}^{0,1}\times\prod_{i\in O(0,1,0)}(1-h_{i1}^{0})\pi_{i1}f_{i1}^{0,0}\times\prod_{i\in O(0,1,1)}h_{i1}^{0}\pi_{i1}f_{i1}^{0,1}\times \\
& \prod_{i\in O(1,0,0)}(1-h_{i0}^{1})\pi_{i0}f_{i0}^{1,0}\times\prod_{i\in O(1,0,1)}h_{i0}^{1}\pi_{i0}f_{i0}^{1,1}\times\prod_{i\in O(1,1,0)}(1-h_{i1}^{1})\pi_{i1}f_{i1}^{1,0}\times\prod_{i\in O(1,1,1)}h_{i1}^{1}\pi_{i1}f_{i1}^{1,1}. \qquad (13)
\end{aligned}
$$

Thus, on the basis of Equations (12) and (13), we can conduct model-based Bayesian inference under SI using two alternative model specifications. Specifically, on the one hand Equation (12) suggests to use a specification similar to that we employ under LSI, specifying (1) the model for principal strata conditional on covariates,

π_{ig}; (2) the model for treatment assigned at time $t = 2$ conditional on intermediate observed outcomes, past treatment and covariates, $h_{iy_1}^{w_1}$; and (3) the model for final potential outcomes conditional on principal strata and covariates, $f_{ig}^{w_1,w_2}$. This specification, which we refer to as specification *SI-1*, implies that the only difference between Bayesian inference under LSI and SI-1 concerns the model for treatment assigned at time $t = 2$, which depends on principal strata when LSI holds, but does only depend on the observed values of the intermediate outcome under SI. On the other hand, on the basis of Equation (13) we can model only the distributions for the observed data, specifying (1) the model for intermediate observed outcomes conditional on covariates, π_{iy_1}; (2) the model for treatment assigned at time $t = 2$ conditional on intermediate observed outcomes, past treatment and covariates, $h_{iy_1}^{w_1}$; and (3) the model for final potential outcomes conditional on on intermediate observed outcomes and covariates, $f_{iy_1}^{w_1,w_2}$. We refer to this model specification based on Equation (13) as specification *SI-2*. Specification SI-2 reflects more closely the standard approaches to causal inference with longitudinal treatments under SI. Specification SI-1 may be preferable if we are interested in the heterogeneity of the effects across principal strata.

6 Simulations

In this Section we investigate the role of LSI (Assumption 4) and sequential ignorability assumptions (Assumption 2 and Assumption 3) using simulations. In our simulated experiment we set up two alternative scenarios in which both the data generating process and the assumptions underlying inference can vary. In the first scenario we generate data under sequential ignorability using a data generating process where both SI and SIL hold, while in the second scenario LSI is assumed. Then we conduct Bayesian inference on the relevant causal estimands for each scenario under both LSI and SI, and in this latter case we use both the SI-1 and the SI-2 specifications. Results from G-methods are also showed when SI/SIL is assumed for comparison purposes.

In order to clearly assess the implications of the two alternative assumptions, LSI and SI/SIL, and investigate the robustness of the estimands to violations of SI/SIL, we focus on a simple setting: we assume that we are in an experimental setting, with no covariates $\boldsymbol{X}_i$ involved, and the first treatment, W_{i1}, randomly assigned, for each unit i. We also consider relatively large sample sizes of 5000 units to avoid (or, at least, reduce) sampling variability issues.

6.1 Data generating processes

The true simulation models for all of the simulations are based on Equation (11) under LSI and Equation (12) under SI, which require to specify parametric models for principal strata (π_{ig}), the treatment assignment probabilities at time $t = 2$ (either $h_{ig}^{w_1}$ or $h_{iy_1}^{w_1}$) and the final outcome ($f_{ig}^{w_1,w_2}$).

The model for principal strata membership contains three conditional probit models, defined using indicator

variables $G_i(11)$, $G_i(00)$ and $G_i(10)$ for whether unit i belongs to principal stratum 11, 00 or 10:

$$\begin{aligned} &G_i(11) = 1 \quad \text{if} \quad G_i^*(11) \equiv \alpha^{11} + \varepsilon_{i,11} \leq 0, \\ &G_i(00) = 1 \quad \text{if} \quad G_i^*(11) > 0 \text{ and } G_i(00)^* \equiv \alpha^{00} + \varepsilon_{i,00} \leq 0, \\ &G_i(10) = 1 \quad \text{if} \quad G_i^*(11) > 0\, G_i(00)^* > 0 \text{ and } G_i(10)^* \equiv \alpha^{10} + \varepsilon_{i,10} \leq 0, \end{aligned} \tag{14}$$

where $\varepsilon_{i,11} \sim N(0,1)$, $\varepsilon_{i,00} \sim N(0,1)$, and $\varepsilon_{i,10} \sim N(0,1)$ independently. Therefore

$$\pi_{i11} = 1 - \Phi\left(\alpha^{11}\right), \qquad \pi_{i00} = \Phi\left(\alpha^{11}\right)\left[1 - \Phi\left(\alpha^{00}\right)\right], \qquad \pi_{i10} = \Phi\left(\alpha^{11}\right)\Phi\left(\alpha^{00}\right)\left[1 - \Phi\left(\alpha^{10}\right)\right]$$

and $\pi_{i01} = 1 - \sum_{g \in \{11,00,10\}} \pi_{ig} = \Phi(\alpha^{11})\Phi(\alpha^{00})\Phi(\alpha^{10})$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard Normal distribution.

For the model of the treatment indicator at time $t = 2$, W_{i2}, we use a probit specification. Under LSI, we assume the following probit model for the treatment assignment at time $t = 2$:

$$W_{i2} = 1 \quad \text{if} \quad W_{i2}^* \equiv \gamma_{w_1} + \gamma_{w_1}^{Y_1(0)} Y_{i1}(0) + \gamma_{w_1}^{Y_1(1)} Y_{i1}(1) + \gamma_{w_1}^{Y_1(0)Y_1(1)} Y_{i1}(0) Y_{i1}(1) + \varepsilon_{i,W_2} > 0, \tag{15}$$

where $\varepsilon_{i,W_2} \sim N(0,1)$, $w_1 \in \{0,1\}$, $Y_{i1}(0) \in \{0,1\}$ and $Y_{i1}(1) \in \{0,1\}$. Under model (15), the treatment assignment probabilities at time $t = 2$ are

$$h_{ig}^{w_1} = \begin{cases} \Phi\left(\gamma_{w_1}\right) & \text{if } W_{i1} = w_1 \text{ and } G_i = 00; \\ \Phi\left(\gamma_{w_1} + \gamma_{w_1}^{Y_1(0)}\right) & \text{if } W_{i1} = w_1 \text{ and } G_i = 10; \\ \Phi\left(\gamma_{w_1} + \gamma_{w_1}^{Y_1(1)}\right) & \text{if } W_{i1} = w_1 \text{ and } G_i = 01; \\ \Phi\left(\gamma_{w_1} + \gamma_{w_1}^{Y_1(0)} + \gamma_{w_1}^{Y_1(1)} + \gamma_{w_1}^{Y_1(0)Y_1(1)}\right) & \text{if } W_{i1} = w_1 \text{ and } G_i = 11; \end{cases}$$

with $w_1 = 0, 1$.

Under SI/SIL, treatment assignment probabilities at time $t = 2$ are free of the missing values for the intermediate outcome, either because W_{i2} does not depends on the missing past potential outcomes or because only observed past potential outcomes enter the assignment mechanism at time $t = 2$. Therefore we impose:

$$\gamma_1^{Y_1(0)} = \gamma_0^{Y_1(1)} = \gamma_0^{Y_1(0)Y_1(1)} = \gamma_1^{Y_1(0)Y_1(1)} = 0 \tag{16}$$

assuming that

$$W_{i2} = 1 \quad \text{if} \quad W_{i2}^* \equiv \gamma_{w_1} + \gamma_{w_1}^{Y_1(w_1)} Y_{i1}(w_1) + \varepsilon_{i,W_2} > 0 \tag{17}$$

where $\varepsilon_{i,W_2} \sim N(0,1)$, $w_1 \in \{0,1\}$, $Y_{i1}(w_1) \in \{0,1\}$. Thus, under SI we have:

$$h_{iy_1}^{w_1} = \begin{cases} \Phi(\gamma_{w_1}) & \text{if } W_{i1} = w_1 \text{ and } y_1 = 0, \\ \Phi\left(\gamma_{w_1} + \gamma_{w_1}^{Y_1(w_1)}\right) & \text{if } W_{i1} = w_1 \text{ and } y_1 = 1. \end{cases}$$

with $w_1 \in \{0,1\}$.

Finally, we need a model for the final outcome, Y_{i2}. In the empirical example the final outcome, Y_{i2}, is the number of employees, which we consider as a continuous variable. Consistently we focus on a continuous final outcome in the simulation studies. Specifically we specify Normal distributions for Y_{i2} conditional on principal strata:

$$\begin{aligned} &Y_{i2}(w_1,w_2)|G_i = g \sim \\ &\quad N\Big(\beta_{w_1w_2} + \beta_{w_1w_2}^{Y_1(0)} Y_{i1}(0) + \beta_{w_1w_2}^{Y_1(1)} Y_{i1}(1) + \beta_{w_1w_2}^{Y_1(0)Y_1(1)} Y_{i1}(0)_{i1}(1); \sigma^2_{w_1w_2,g}\Big), \end{aligned} \tag{18}$$

$w_1 \in \{0,1\}$, $w_2 \in \{0,1\}$, $g \in \{00,01,10,11\}$. For simplicity, we impose prior equality of the variance parameters across principal strata: $\sigma^2_{w_1w_2,g} = \sigma^2_{w_1w_2}$, $g \in \{00,01,10,11\}$. Let $\Phi(\mu;\sigma^2)$ denote the probability density function of a Normal distribution with mean μ and variance σ^2. Then, the Normal distributions in Equation (18) implies that:

$$f_{ig}^{w_1,w_2} = \begin{cases} f\left(\beta_{w_1w_2}; \sigma^2_{w_1w_2}\right) & \text{if } W_{i1} = w_1, W_{i2} = w_2, \text{ and } G_i = 00; \\ f\left(\beta_{w_1w_2} + \beta_{w_1w_2}^{Y_1(0)}; \sigma^2_{w_1w_2}\right) & \text{if } W_{i1} = w_1, W_{i2} = w_2, \text{ and } G_i = 10; \\ f\left(\beta_{w_1w_2} + \beta_{w_1w_2}^{Y_1(1)}; \sigma^2_{w_1w_2}\right) & \text{if } W_{i1} = w_1, W_{i2} = w_2, \text{ and } G_i = 01; \\ f\left(\beta_{w_1w_2} + \beta_{w_1w_2}^{Y_1(0)} + \beta_{w_1w_2}^{Y_1(1)} + \beta_{w_1w_2}^{Y_1(0)Y_1(1)}; \sigma^2_{w_1w_2}\right) & \text{if } W_{i1} = w_1, W_{i2} = w_2, \text{ and } G_i = 11; \end{cases}$$

$w_1 \in \{0,1\}$, $w_2 \in \{0,1\}$, where $f(\mu,\sigma^2)$ is the probability density function of a normal distribution with mean μ and variance σ^2. We use this model specification to generate data for the final outcome under both LSI and SI. Therefore, the complete parameter vector for the simulation models is $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$, where $\boldsymbol{\alpha} = (\alpha^{11}, \alpha^{00}, \alpha^{10})$, $\boldsymbol{\gamma} = \{\gamma_{w_1}, \gamma_{w_1}^{Y_1(0)}, \gamma_{w_1}^{Y_1(1)}, \gamma_{w_1}^{Y_1(0)Y_1(1)}\}_{w_1 \in \{0,1\}}$, $\boldsymbol{\beta} = \{\beta_{w_1w_2}, \beta_{w_1w_2}^{Y_1(0)}, \beta_{w_1w_2}^{Y_1(1)}, \beta_{w_1w_2}^{Y_1(0)Y_1(1)}\}_{w_1 \in \{0,1\}, w_2 \in \{0,1\}}$ and $\boldsymbol{\sigma}^2 = \{\sigma^2_{w_1w_2}\}_{w_1 \in \{0,1\}, w_2 \in \{0,1\}}$. The parameter vector $\boldsymbol{\theta}$ includes 31 parameters, 4 of which are forced to be equal to 0 when SI/SIL holds according to Equation (16). The true values of these parameters are given in the Supplementary Materials available on-line.

6.2 Inference in the simulations

For each simulation scenario we conduct inference under both LSI and SI/SIL. Under LSI we assume the parametric models specified in Equations (14), (15) and (18) in Section 6.1. It is worth noting that inference under LSI always uses a correct model specification, even when SI holds. In fact, when SI/SIL holds, some parameters, namely $\gamma_1^{Y_1(0)}$, $\gamma_0^{Y_1(1)}$, $\gamma_0^{Y_1(0)Y_1(1)}$ and $\gamma_1^{Y_1(0)Y_1(1)}$, are simply equal to zero.

Under the specification of type SI-1 we assume the parametric models specified in Equations (14), (17)

and (18) in Section 6.1. Under the specification of type SI-2 we specify the probit model in Equation (17) for treatment assignment at time $t = 2$, and the following models for the intermediate and final outcomes. We use a probit specification for the intermediate outcome, $Y_{i1}(w_1)$:

$$Y_{i1}(w_1) = 1 \quad \text{if} \quad Y^*_{i1}(w_1) \equiv \alpha_{w_1} + \varepsilon_{i,Y_1(w_1)} > 0 \tag{19}$$

where $\varepsilon_{i,Y_1(w_1)} \sim N(0,1)$, $w_1 \in \{0,1\}$, and we posit a Normal model on the final outcome $Y_{i2}(w_1,w_2)$, conditional on $Y_1(w_1)$:

$$Y_{i2}(w_1,w_2)|Y_{i1}(w_1) = y_1 \sim N\Big(\beta_{w_1w_2} + \beta^{Y_1(w_1)}_{w_1w_2} y_1;\, \sigma^2_{w_1w_2,y_1}\Big) \tag{20}$$

$w_1 \in \{0,1\}$, $w_2 \in \{0,1\}$, $y_1 \in \{0,1\}$. Again we impose $\sigma^2_{w_1w_2,y_1=0} = \sigma^2_{w_1w_2,y_1=1} \equiv \sigma^2_{w_1w_2}$. We assume that parameters are a priori independent and use proper, although non informative, Normal (or Scale-Inv-χ^2 for $\sigma^2_{w_1w_2}$) prior distributions.

Posterior inference on $\boldsymbol{\theta}$ is obtained using Markov Chain Monte Carlo (MCMC) methods. The MCMC algorithms we adopted under LSI and SI with specification SI-1 use Gibbs sampler with data augmentation to impute at each step the missing principal stratum membership, G_i. Under SI with specification SI-2, the likelihood function does not involve mixtures of distributions associated with the latent strata, but only depends on observed distributions, so the posterior distribution of $\boldsymbol{\theta}$ can be easily derived using Gibbs sampling methods. See on-line Supplementary Materials for further details on the prior distributions and the MCMC algorithms. The posterior distributions were simulated running a chain for 9000 MCMC iterations, after an initial 1000 burn-in iterations.

Results obtained using saturated marginal structural models, estimated by means of inverse probability of treatment weighting in a frequentist fashion, are show in the Supplementary Materials available on-line.

6.3 Simulation Results

Simulation results for the causal estimands of interest are shown in Tables 2 and 3 and in Figures 1 and 2. Table 2 and Table 3 show posterior means, standard deviations and 95% posterior credible intervals for the average causal effects in Equation (1) when LSI and SI/SIL holds, respectively, and inference is conducted using LSI, SI-1 and SI-2 specifications. Similarly, under the same scenarios and model specifications, Figures 1 and 2 depict the posterior distributions of the six average causal effects.

Table 2 and Figure 1 make it clear that when LSI is the true assumption behind the data generating process, inference under sequential ignorability assumptions may lead to misleading results. Figure 1 shows that only the LSI inferential framework is able to always lead to valid inference about the six causal effects of interest. All the posterior distributions of the six ATEs derived under LSI reach their maximum in a thin neighbourhood around the true ATE values, so the posterior modes, which approximately correspond to the posterior means, appear to be good point estimates for the causal effects of interest. Also the posterior variability is relatively small and the 95% posterior credible intervals, which always cover the true ATE values, are quite narrow, making inference very precise. Conversely, assuming SI/SIL, when LSI actually holds, may yield to completely wrong inferences, especially when a specification of type SI-2 is used. The posterior distributions of the six ATEs derived under SI with a specification of type SI-1 cover the true ATE values only in the queue for most of the six causal effects.

Table 2: Summary statistics for the causal estimands posterior distributions when LSI holds.

		LSI				SI-1				SI-2			
Estimand	true	mean	sd	2.5%	97.5%	mean	sd	2.5%	97.5%	mean	sd	2.5%	97.5%
$ATE_{11.00}$	12.54	12.41	0.19	11.91	12.76	12.17	0.17	11.74	12.52	12.21	0.19	11.84	12.58
$ATE_{11.01}$	6.25	6.17	0.32	5.50	6.92	5.56	0.26	5.01	6.15	2.24	0.19	1.87	2.61
$ATE_{11.10}$	7.54	7.52	0.25	6.96	7.99	6.97	0.23	6.43	7.46	3.64	0.18	3.27	4.00
$ATE_{10.00}$	5.01	4.89	0.20	4.61	5.26	5.21	0.18	4.93	5.51	8.57	0.16	8.25	8.88
$ATE_{01.10}$	1.29	1.33	0.32	0.44	2.14	1.41	0.28	0.79	2.05	1.40	0.16	1.09	1.71
$ATE_{01.00}$	6.29	6.22	0.27	5.57	6.77	6.61	0.22	6.21	6.96	9.97	0.16	9.64	10.28

Specifically, the 95% posterior credible intervals cover the true ATE values only for three out of the six ATEs: $ATE_{10.00}$, $ATE_{01.10}$, and $ATE_{01.00}$. For the remain causal estimands, $ATE_{11.00}$, $ATE_{11.01}$, and $ATE_{11.10}$, the true values are extreme values according to the estimated posterior distributions. The handicaps of inference under sequential ignorability assumptions is even more dramatic when a specification of type SI-2 is used. In such a case, the posterior distributions for four out of six ATEs are located far away from the true values.

The different performances of inference under SI/SIL, comparing specifications of type SI-1 and SI-2, may be (at least partially) justified noting that the first specification accounts for heterogeneity in the distribution of the final outcome, and thus in the causal effects, across principal strata. Therefore, although focus is on average causal effects for the whole population, if principal strata are strongly associated with the final outcomes, a parametrization of type SI-1 may in some sense address the consequences of a misspecified treatment assignment mechanism.

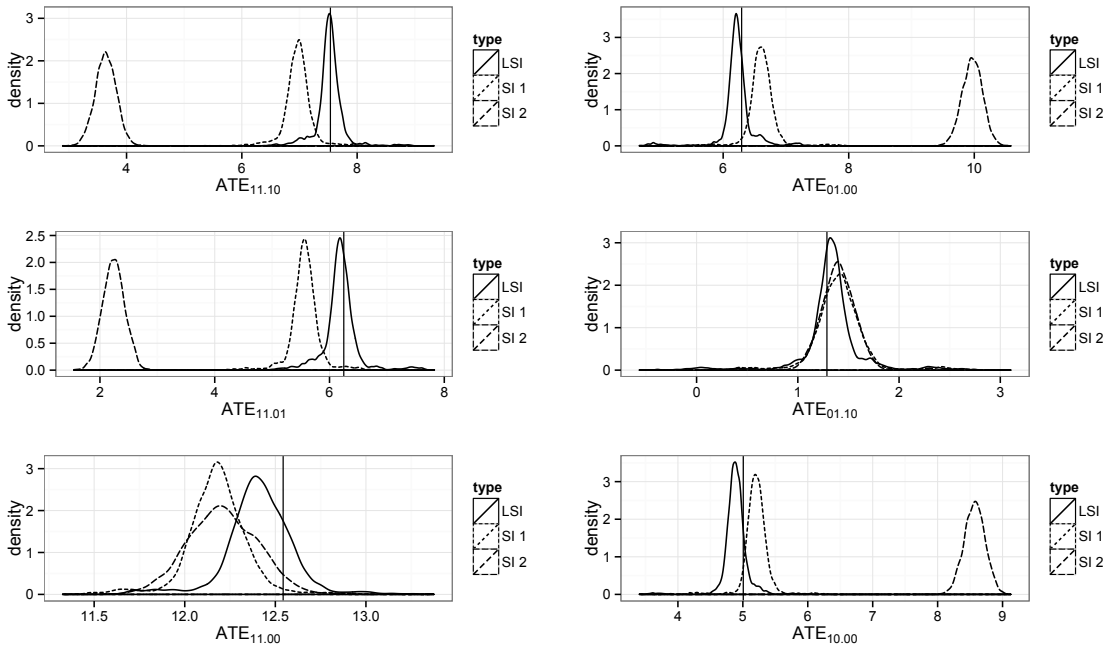

Figure 1: Posterior density functions for the ATEs under the LSI scenario. Inference using LSI (solid), SI-1 (dotted) and SI-2 (dashed). The vertical solid line indicates the true value of the ATE.

When SI/SIL is the true assumption underlying the data generation process, both LSI and SI/SIL (with either a specification of type SI-1 or a specification of type SI-2) lead to valid inferences about the causal estimands of interest. As can be seen in Figure 2 and Table 3, the posterior distributions are bell-shaped approximately

Table 3: Summary statistics for the causal estimands posterior distributions when SI/SIL holds.

		LSI				SI-1				SI-2			
Estimand	true	mean	sd	2.5%	97.5%	mean	sd	2.5%	97.5%	mean	sd	2.5%	97.5%
$ATE_{11.00}$	12.54	12.52	0.25	12.00	12.98	12.62	0.26	12.13	13.08	12.42	0.18	12.06	17.77
$ATE_{11.01}$	6.25	6.16	0.31	5.51	6.91	6.31	0.31	5.60	7.00	6.14	0.21	5.73	6.55
$ATE_{11.10}$	7.54	7.49	0.27	6.85	8.12	7.56	0.28	6.88	8.22	7.53	0.19	7.14	7.90
$ATE_{10.00}$	5.01	5.02	0.17	4.65	5.31	5.06	0.17	4.68	5.35	4.89	0.18	4.55	5.23
$ATE_{01.10}$	1.29	1.33	0.25	0.62	1.90	1.24	0.24	0.66	1.85	1.39	0.21	0.99	1.79
$ATE_{01.00}$	6.29	6.35	0.23	5.66	6.84	6.30	0.21	5.78	6.80	6.28	0.18	5.92	6.63

symmetrical curves around the true ATE values, providing relatively narrow 95% posterior credible intervals, which always cover the true ATE values. These results suggest that comparing inferences about the causal effects of interest derived under LSI and SI/SIL may provide useful insights on the plausibility of sequential ignorability assumptions. Further evidence can be obtained looking at the posterior distributions of the treatment assignment probabilities at time $t = 2$, given the treatment received at time $t = 1$ and principal stratum membership, derived under the LSI assumption, and investigate if equalities in Equation (6) may hold. Equalities in Equation (6) are indeed key quantities to assess the plausibility of sequential ignorability assumptions. LSI is a relaxed version of SIL: SIL can be viewed as a special case of LSI, where the equalities in Equation (6) hold. Therefore if equalities in Equation (6) do not hold, SIL clearly does not hold either. On another hand, if equalities in Equation (6) do not hold, doubts on the plausibility of SI (Assumption 2) also arise, because we can reasonably expect that the assignment probabilities at time $t = 2$ depend on unobserved characteristics related to the final outcome, which make SI untenable.

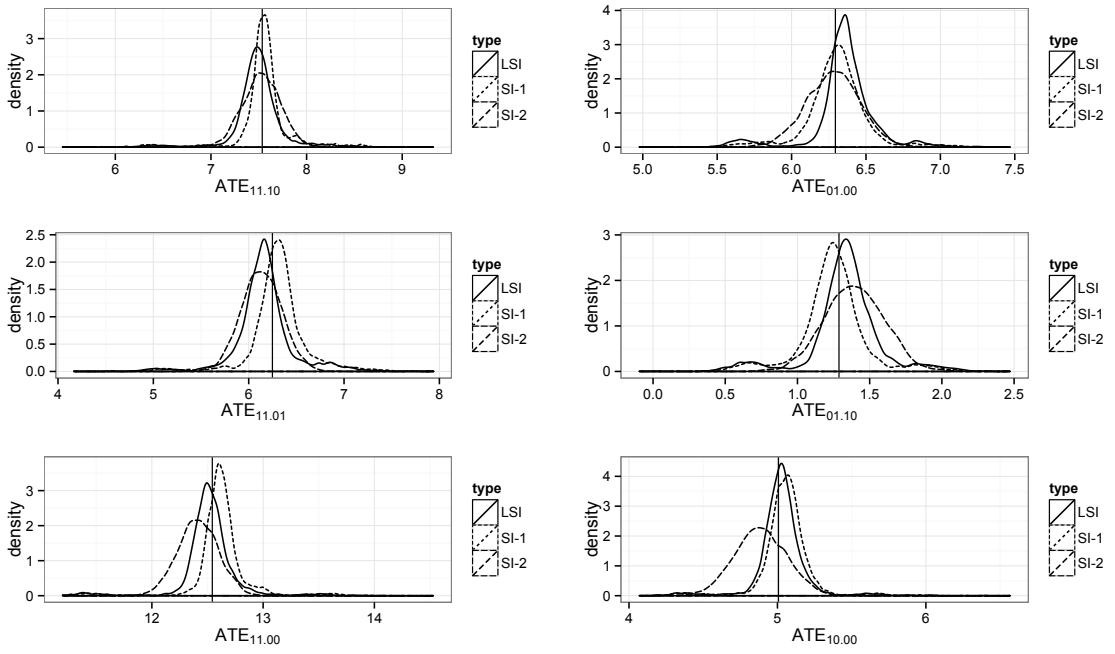

Figure 2: Posterior density functions for the ATEs under the SI scenario. Inference using LSI (solid), SI-1 (dotted) and SI-2 (dashed). The vertical solid line indicates the true value of the ATE.

Figures 3 and 4 show the posterior distributions of those treatment assignment probabilities coupled according to the equalities in Equation (6). As we could expect, the posterior distributions of the treatment assignment

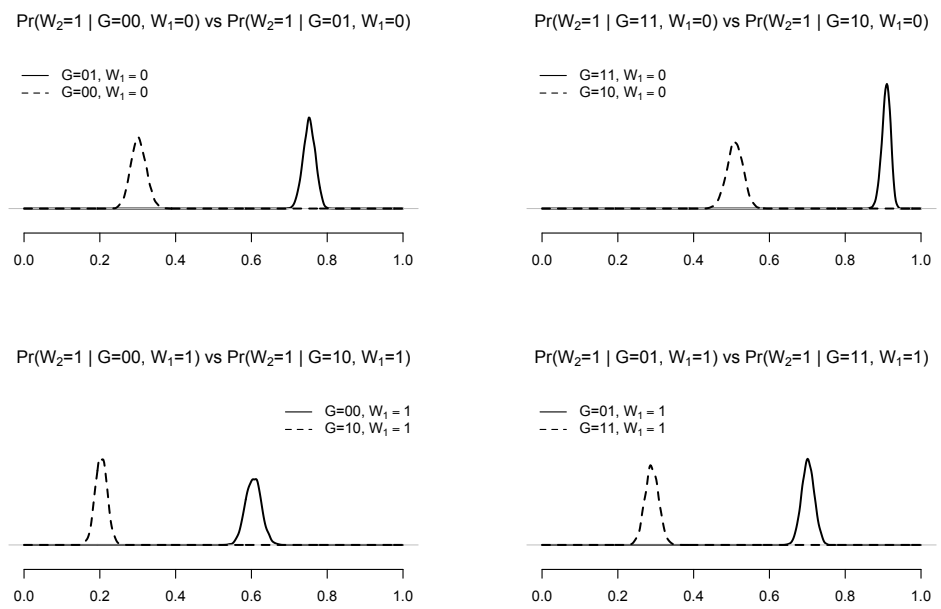

Figure 3: Posterior density functions for the assignment probabilities when LSI holds.

probabilities at time $t = 2$ in Figure 3 appear to be highly heterogeneous across principal strata, suggesting strong evidence against the equalities in Equation (6), and thus, against sequential ignorability. On the other hand, the posterior distributions in Figure 4 look very similar and are highly overlapping, showing no evidence against sequential ignorability in favor of LSI.

It is worth noting that our simulation results are based on a single simulated data set. Studying the frequentist properties of our procedure in repeated samples is beyond the scope of the paper.

7 An illustrative example: the Tuscan Government Founding Program

We illustrate our framework in a program evaluation study concerning causal effects on firms' performances of an interest free loans policy aiming to ease access to credit by making it less costly. Firms meeting certain standards to be eligible can apply to get an interest free loan at various points in time, thus firms may apply and be granted multiple times over subsequent years.

The program started in 2002 and was rolling on a yearly basis. In this paper we consider data in the years between 2002 and 2007, and we focus on casual effects defined by contrasting firms' performances measured in terms of employment levels at the end of the study, in 2007, under different treatment sequences. Treatment sequences are defined using two binary treatment variables: W_{i1} equal to one if firm i is granted at least one time between 2002 and 2004, and zero otherwise; and W_{i2} equal to one if firm i is granted at least one time between 2005 and 2007, and zero otherwise. The final outcome of interest, Y_{i2}, is the firm i's number of employees at the end of 2007. As intermediate outcome we consider a binary variable Y_{i1} describing the hiring policy of firm i at the end of 2004: $Y_{i1} = 1$ if firm i hires new employees by the end of 2004, and $Y_{i1} = 0$ otherwise. Therefore the basic

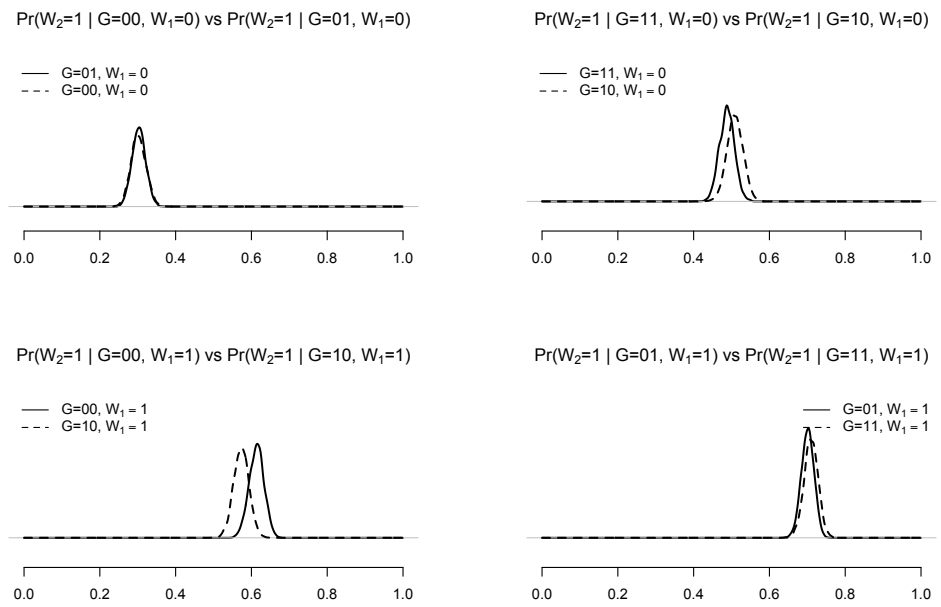

Figure 4: Posterior density functions for the assignment probabilities when SI holds.

principal stratification with respect to this intermediate variable classifies firms in four latent groups: $G_i = 00$ comprising firms that would not hire irrespective of the treatment received at time $t = 1$; $G_i = 01$ comprising firms that would hire if granted but would not hire if not granted; $G_i = 10$ comprising firms that would hire if not granted and would hire if granted; and $G_i = 11$ comprising firms that would always hire irrespective of the treatment received at time $t = 1$. Principal strata with respect to the indicator for hiring choices can be viewed as a coarsened representation of the latent hiring preferences of a firm, which are reasonably associated with both the decision to participate in the treatment at a subsequent time point and the final outcome.

Firms exposed to different treatment sequences are likely to differ in many dimensions. Thus, preliminary analyses based on matching procedures were conducted to create a sub-population of firms exposed to alternative treatment sequences, where the distributions of the confounders, including both pretreatment baseline covariates and intermediate (time-variant) confounder variables, are overlapping and well balanced across firms exposed to alternative treatment sequences (see Pirani et al. 2013, for details). For this reason we did not use any covariate in our models for the inference. Our final sample consists of 4615 firms, among which 632 firms received a loan only in one treatment period, 33 firms received a loan in both treatment periods, and the rest did not received any loan in the observation period.

Bayesian inference for the average casual effects of interest was conducted under both LSI and SI/SIL. Under LSI the model we specified involved the tree sub-models described in Equations (14), (15) and (18). Bayesian inference under SI/SIL was conducted using both a specification of type SI-1, which involved the tree sub-models described in Equations (14), (17) and (18), and a specification of type SI-2, which involved the tree sub-models described in Equations (19), (17) and (20).

Table 4 shows the posterior means, standard deviations and 95% posterior credible intervals for the stratum

Table 4: Summary statistics for the causal estimands posterior distributions in the real case.

	LSI				SI-1				SI-2			
Estimand	mean	sd	2.5%	97.5%	mean	sd	2.5%	97.5%	mean	sd	2.5%	97.5%
$\widehat{\pi}_{00}$	0.69	0.01	0.67	0.70	0.69	0.01	0.67	0.70	–	–	–	–
$\widehat{\pi}_{01}$	0.04	0.00	0.03	0.05	0.04	0.00	0.03	0.04	–	–	–	–
$\widehat{\pi}_{10}$	0.23	0.01	0.22	0.25	0.24	0.01	0.22	0.25	–	–	–	–
$\widehat{\pi}_{11}$	0.04	0.00	0.03	0.05	0.04	0.00	0.03	0.05	–	–	–	–
$ATE_{11.00}$	15.11	6.03	-0.69	21.62	8.22	3.47	1.63	15.70	6.88	2.96	0.90	12.66
$ATE_{11.01}$	10.19	6.07	-5.69	16.83	3.57	3.53	-3.15	11.20	3.87	3.01	-2.16	9.58
$ATE_{11.10}$	11.73	6.01	-3.97	18.36	5.02	3.50	-1.60	12.52	5.00	3.00	-1.06	10.85
$ATE_{10.00}$	3.39	0.51	2.38	4.39	3.20	0.51	2.21	4.18	1.88	0.52	0.89	2.88
$ATE_{01.10}$	1.54	0.70	0.15	2.93	1.45	0.70	0.09	2.80	1.13	0.77	-0.39	2.59
$ATE_{01.00}$	4.92	0.55	3.85	6.07	4.65	0.56	3.52	5.71	3.01	0.58	1.90	4.13

membership probabilities and the six ATEs, while Figure 5 portrays their posterior density functions. Inference under SI/SIL does not seem to strongly depend on the type of specification used. Both specifications SI-1 and SI-2 lead to similar results for all the six causal effects of interest but two, $ATE_{01.00}$ and $ATE_{10.00}$. The posterior means of $ATE_{01.00}$ and $ATE_{10.00}$ derived under specification SI-1 are greater than those derived under specification SI-2.

Under LSI we obtain substantially different inferential results than those under SI, especially for some causal estimands. LSI leads to posterior distributions for the causal effects $ATE_{10.00}$, $ATE_{11.01}$ and $ATE_{11.10}$ that are essentially the same as those derived under SI/SIL with a specification of type SI-1, but provides posterior distributions for $ATE_{11.00}$, $ATE_{11.01}$ and $ATE_{11.10}$ that are centered on much higher values and have a higher variability than those derived under SI/SIL. Specifically the posterior means of $ATE_{11.00}$, $ATE_{11.01}$ and $ATE_{11.10}$ derived under LSI are more than 1.8 times those derived under SI/SIL, but the posterior standard deviations of $ATE_{11.00}$, $ATE_{11.01}$ and $ATE_{11.10}$ derived under LSI are about twice those derived under SI/SIL. The loss of precision we have assuming LSI rather than SI/SIL is probably due to the greater complexity of the model, which is only weakly identified (Gustafson 2010); the small proportion of firms that is estimated to belong to some principal strata (e.g., the posterior means of the probabilities to belongs to groups 01 and 11 are equal to 0.04, see $\widehat{\pi}_g$ in Table 4); and the extremely low estimated probability to receive a loan at time $t = 2$ for all firms but those in the principal stratum 01 that received a loan at time $t = 1$ (see Table 5).

Given that LSI and SI/SIL lead to quite different posterior distributions of the causal estimands of interest, it becomes of compelling interest to investigate the plausibility of the sequential ignorability assumptions. To this end, we look at the posterior distributions of the assignment probabilities at time $t = 2$, summarized in Table 5. As we can see,the probability to receive a loan in the second treatment period is very small for all firms expect for those belonging to principal stratum 01 that received a loan in the first treatment period. Also we find large differences comparing the posterior distributions of the assignment probabilities at time $t = 2$ two by two according to the equalities in Equation (6). Among firms that did not receive a loan in the first treatment period ($W_{i1} = 0$) the posterior mean of the probability to receive a loan in the second treatment period is about 7% for firms in principal stratum 10 and about 1% for firms in principal stratum 11. Among firms that received a loan in the first treatment period ($W_{i1} = 1$) the posterior mean of the probability to receive a loan in the second treatment period is about 76% for firms in principal stratum 01 and less than $5‰$ for firms in principal stratum 11. The posterior distributions of the probability to receive a loan in the second treatment period for firms in principal strata 00 and 01 assigned to $W_{i1} = 0$ and for firms in principal strata 00 and 10 assigned to $W_{i1} = 1$ are more similar, although there is still strong evidence again the assumption that they are the same, as they should be under SIL. Therefore

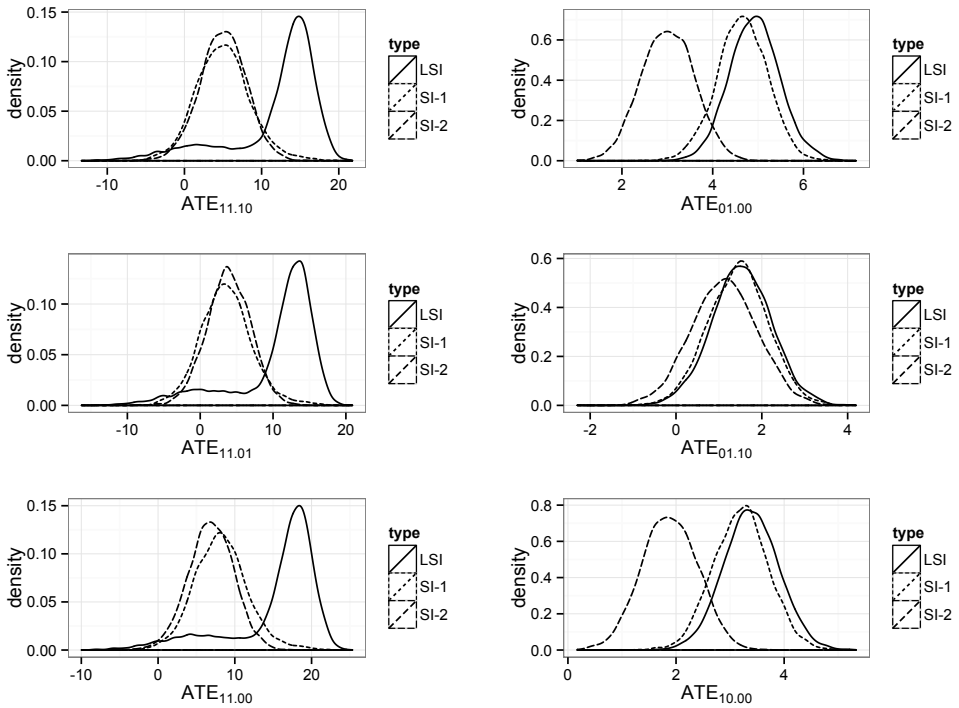

Figure 5: Posterior density functions for the ATEs in the real case application. Inference using LSI (solid), SI-1 (dotted) and SI-2 (dashed).

Table 5: Summary statistics for the assignment probabilities in the real case.

W_1	G	mean	sd	Q_1	median	Q_3
0	00	0.0690	0.0048	0.0657	0.0689	0.0721
	01	0.0432	0.0170	0.0309	0.0414	0.0532
	10	0.0708	0.0081	0.0652	0.0706	0.0762
	11	0.0105	0.0100	0.0032	0.0074	0.0149
1	00	0.0890	0.0418	0.0795	0.0990	0.1156
	10	0.0765	0.1167	0.0139	0.0267	0.0538
	01	0.7627	0.1326	0.6785	0.7812	0.8643
	11	0.0004	0.0018	0.0000	0.0000	0.0000

our results show some evidence that sequential ignorability assumptions are questionable in this study, suggesting inference under LSI to be more reliable here.

8 Concluding Remarks

We focus on the role of the critical assumptions about the assignment mechanism in causal inference for time-varying treatments, proposing a new assumption, that we call latent sequential ignorability (LSI), which may be more reasonable than the usually invoked sequential ignorability assumptions in some settings. LSI implies that the joint values of potential outcomes for the relevant intermediate variables (i.e., the principal strata), rather than their observed values only, include crucial information about the decision to participate in the treatment. Therefore LSI focuses on specific violations of SI due to the presence of unobserved factors affecting the decision to participate in the treatment that can be summarized by principal strata.

In studies where ignorability assumptions are not reasonable, LSI provides a powerful framework, which also

permits to easily assess the sensitivity of inferential conclusions with respect to violations of sequential ignorability assumptions (SI and SIL) implied by LSI looking at inferences on the probabilities of treatment assignment at a given time point under LSI, conditional on the observed history and principal strata. These quantities are key estimands in causal inference under LSI, so no additional effort is required to perform sensitivity analysis under LSI.

Simulation results show that LSI conducts to valid inference for causal effects even if SI/SIL holds, although it usually involves more complex models. On the other hand, inference under sequential ignorability assumptions may lead to very misleading inferential conclusions when it does not hold, but LSI does.

In our illustrative example, sensitivity analysis showed strong evidence against sequential ignorability assumptions: the posterior distributions of the treatment assignment probabilities at time $t = 2$ withing groups defined by the first treatment and principal stratum membership are quite heterogeneous, suggesting that inferences based on LSI are more reliable.

Another appealing feature of LSI is that it provides a natural framework to investigate the heterogeneity of the effects across principal strata. Assessing causal effects stratified by intermediate outcomes under SI/SIL generally requires additional efforts. In particular, in a Bayesian setting, one needs to specify a model for principal strata membership conditional on the covariates, and a model for the potential outcomes $Y_{i2}(w_1, w_2)$ conditional on principal strata and covariates. This model specification, which corresponds to using a specification of type SI-1 under SI/SIL, is the core of the inferential approach under LSI, but it is not standard in causal inference under sequential ignorability assumptions. Here we did not investigate issues concerning causal effect heterogeneity across principal strata, focusing on comparing inferences about causal effects for the whole population under LSI and sequential ignorability assumptions, however the heterogeneity of the effects with respect to principal strata can be often of interest to policy makers.

As a general message, our study stresses the importance of carefully evaluating the plausibility of the assumptions underlying the analysis, especially in complex settings like those arising in longitudinal observational studies.

The extension of our framework to multiple (i.e., more than two) time points is without any doubt an interesting future development. The extension to additional periods is conceptually straightforward, but raises challenging practical issues due to the fact that the number of principal strata increases exponentially with the number of time points. To cope with the increasingly huge missing data problem, additional assumptions are required. For instance we could invoke Markovian properties, similarly to what has been done, e.g., both by Lin et al. (2008) in a study in which units are randomized at the baseline and compliance to treatment may vary longitudinally, and by Dai et al. (2012) who considered, again, a unique treatment and some post-randomization time-varying behavioural variables.

Acknowledgements

Financial support for this research was provided through the grant "Programma Futuro in Ricerca 2012 – RBFR12SHVV_003", financed by the Italian Ministero dell'Istruzione, dell'Università e della Ricerca.

References

Achy-Brou, A. C., Frangakis, C. E. & Griswold, M. (2010), 'Estimating treatment effects of longitudinal designs using regression models on propensity scores', *Biometrics* **66**, 824–833.

Dai, J. Y., Gilbert, P. B. & Mâsse, B. R. (2012), 'Partially hidden markov model for time-varying principal stratification in hiv prevention trials.', *Journal of the American Statistical Association* **107**, 52–65.

de Finetti, B. (1937), La prévision: ses lois logiques, ses sources subjectives, *in* 'Annales de l'institut Henri Poincaré', Vol. 7, Presses universitaires de France, pp. 1–68.

de Finetti, B. (1964), Foresight: Its logical laws, its subjective sources., *in* H. E. Kyburg & H. E. Smokler, eds, 'Studies in Subjective Probability', New York: Wiley, pp. 93–158.

Ding, P. & VanderWeele, T. J. (2016), 'Sensitivity analysis without assumptions', *Epidemiology* **27**, 368–377.

Frangakis, C. E. & Rubin, D. B. (1999), 'Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes.', *Biometrika* **86**, 365–379.

Frangakis, C. E. & Rubin, D. B. (2002), 'Principal stratification in causal inference.', *Biometrics* **58**(1), 21–29.

Gill, R. D. & Robins, J. M. (2001), 'Causal inference for complex longitudinal data: The continuous case', *The Annals of Statistics* **29**, 1785–1811.

Gustafson, P. (2010), 'Bayesian inference for partially identified models', *The International Journal of Biostatistics* **6**(2), Article 17.

Heckman, J. J. & Navarro, S. (2007), 'Dynamic discrete choice and dynamic treatment effects', *Journal of Econometrics* **136**, 341–396.

Holland, P. (1986), 'Statistics and Causal Inference.', *Journal of the American Statistical Association* **81**, 945–960.

Hong, G. & Raudenbush, S. W. (2008), 'Causal inference for time-varying instructional treatments', *Journal of Educational and Behavioral Statistics* **33**, 333–362.

Ichino, A., Mealli, F. & Nannicini, T. (2008), 'From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity?', *Journal of Applied Econometrics* **23**, 305–327.

Imai, K. & Ratkovic, M. (2014), 'Robust estimation of inverse probability weights for marginal structural models', *Journal of the American Statistical Association. To apper* .

Imbens, G. & Rubin, D. B. (1997), 'Bayesian inference for causal effects in randomized experiments with noncompliance.', *The Annals of Statistics* **25**, 305–327.

Imbens, G. & Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Imbens, G. W. (2003), 'Sensitivity to exogeneity assumptions in program evaluation', *The American Economic Review,* **93**, 126–132.

Lechner, M. (2009), 'Sequential causal models for the evaluation of labor market programs', *Journal of Business and Economic Statistics* **27**, 71–83.

Lin, J. Y., Ten Have, T. & Elliott, M. R. (2008), 'Longitudinal Nested Compliance Class Model in the Presence of Time-Varying Noncompliance.', *Journal of the American Statistical Association* **103**, 505–513.

Mattei, A. & Mealli, F. (2007), 'Application of the principal stratification approach to the faenza randomized experiment on breast self-examination', *Biometrics* pp. 437–446.

Murphy, S. A. (2003), 'Optimal dynamic treatment regimes', *Journal of the Royal Statistical Society, Series B* **65**, 331–355.

Pirani, E., Mariani, M. & Mealli, F. (2013), A longitudinal design for the evaluation of the effects of subsidized loans on firm survival and growth. presented at the 1st *Southern European Conference on Survey Methodology*, 12-14 december 2013, barcellona, spain.

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

Robins, J. M. (1989), 'The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitu- dinal studies', *Health Service Research Methodology: A Focus on AIDS* pp. 113–159.

Robins, J. M. (1997), 'Causal inference from complex longitudinal data', *Latent Variable Modeling and Applications to Causality* **120**, 69–117.

Robins, J. M. (1999), Marginal structural models versus structural nested models as tools for causal inference, *in* M. Halloran & D. Berry, eds, 'Statistical Models in Epidemiology: The Environment and Clinical Trials', New York: Springer-Verlag, pp. 95–134.

Robins, J. M. (2004), Optimal structural nested models for optimal sequential decisions, *in* 'Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data', pp. 189–326.

Robins, J. M., Hernán, M. A. & Brumback, B. (2000), 'Marginal structural models and causal inference', *Epidemiology* **11**, 550–560.

Rosenbaum, P. (1987), 'Sensitivity analysis to certain permutation inferences in matched observational studies', *Biometrika* **74**, 13–26.

Rosenbaum, P. R. & Rubin, D. B. (1983), 'Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome', *Journal of the Royal Statistical Society, Series B: Methodological* **45**, 212–218.

Rubin, D. B. (1974), 'Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies', *J. Educ. Psychol.* **66**(5), 688–701.

Rubin, D. B. (1977), 'Assignment to Treatment Group on the Basis of a Covariate.', *Journal of Educational Statistics* **2**, 1–26.

Rubin, D. B. (1978), 'Bayesian Inference for Causal Effects: The Role of Randomization.', *The Annals of Statistics* **6**, 34–58.

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

Rubin, D. B. (2005), 'Causal Inference Using Potential Outcomes: Design, Model- ing, Decisions', *Journal of the American Statistical Association* **100**(469), 322–331.

Schwartz, S. L., F., L. & F., M. (2011), ' A Bayesian semiparametric approach to intermediate variables in Causal inference', *Journal of the American Statistical Association* **106**, 1331–1344.

Zajonc, T. (2012), 'Bayesian inference for dynamic treatment regimes: Mobility, equity, and efficiency in student tracking', *Journal of the American Statistical Association* **107**(497), 80–92.