Coherent modeling of longitudinal causal effects on binary outcomes

Linbo Wang

Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5S 3G3, Canada

email: linbo.wang@utoronto.ca

and

Xiang Meng

Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A.

email: xmeng@g.harvard.edu

and

Thomas S. Richardson

Department of Statistics, University of Washington, Seattle, Washington 98195, U.S.A.

email: thomasr@u.washington.edu

and

James M. Robins

Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.

email: robins@hsph.harvard.edu

SUMMARY: Analyses of biomedical studies often necessitate modeling longitudinal causal effects. The current focus on personalized medicine and effect heterogeneity makes this task even more challenging. Towards this end, structural nested mean models (SNMMs) are fundamental tools for studying heterogeneous treatment effects in longitudinal studies. However, when outcomes are binary, current methods for estimating multiplicative and additive SNMM parameters suffer from variation dependence between the causal parameters and the non-causal nuisance parameters. This leads to a series of difficulties in interpretation, estimation and computation. These difficulties have hindered the uptake of SNMMs in biomedical practice, where binary outcomes are very common. We solve the variation dependence problem for the binary multiplicative SNMM via a reparametrization of the non-causal

BIOMETRICS 63, 000–000 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2007

nuisance parameters. Our novel nuisance parameters are variation independent of the causal parameters, and hence allow for coherent modeling of heterogeneous effects from longitudinal studies with binary outcomes. Our parametrization also provides a key building block for flexible doubly robust estimation of the causal parameters. Along the way, we prove that an additive SNMM with binary outcomes does not admit a variation independent parametrization, thereby justifying the restriction to multiplicative SNMMs.

KEY WORDS: Bivariate mapping; Likelihood inference; Longitudinal studies; Variation independence

1. Introduction

In biomedical studies researchers are often interested in inferring causal effects from longitudinal studies with time-dependent exposures. For example, suppose one is interested in estimating the (joint) effect of maternal stress on childhood illness from longitudinal observational data. The relationships among observed variables may be represented by the causal directed acyclic graph (DAG, Pearl, 2009) in Figure 1, in which A_0 and A_1 denote maternal stress levels at baseline and the first follow-up, respectively, L_1 is the intermediate covariate encoding whether or not the child is ill at the first follow-up, and Y is the outcome of interest encoding whether or not the child is ill at the second follow-up. The node U denotes unmeasured variables such as the child's underlying immune status. There may also be covariates L_0 measured at baseline, in which case one can add L_0 and a directed edge from L_0 to every other node in Figure 1.

[Figure 1 about here.]

As pointed out by Robins (1986), conventional regression adjustment methods cannot be used to estimate the joint effect of A_0 and A_1, regardless of whether or not one adjusts for the time-dependent confounder L_1 in the regression. Instead, Robins (1986) proposed the so-called g-formula

$$E[Y(a_0, a_1)] = \sum_{l_0, l_1} E[Y \mid l_0, a_0, l_1, a_1] f(l_1 \mid l_0, a_0) f(l_0), \tag{1}$$

where $Y(a_0, a_1)$ denotes the potential outcome, $E[Y \mid l_0, a_0, l_1, a_1] = E[Y \mid L_0 = l_0, A_0 = a_0, L_1 = l_1, A_1 = a_1]$, $f(l_1 \mid l_0, a_0)$ and $f(l_0)$ represent the (conditional) density of L_1 and L_0, respectively. Application of the g-formula in practice, however, is subject to the g-null paradox (Robins, 1986; Robins and Wasserman, 1997). Specifically, under the causal null hypothesis represented by Figure 2, in general, both $E[Y \mid l_0, a_0, l_1, a_1]$ and $f(l_1 \mid l_0, a_0)$ depend on a_0. In this case, it is very challenging to specify models for $E[Y \mid l_0, a_0, l_1, a_1]$ and $f(l_1 \mid l_0, a_0)$ that allow for each term to depend on a_0 while at the same time being compatible with the causal null hypothesis that (1) does not depend on a_0.

[Figure 2 about here.]

In response to the g-null paradox, Robins (1994) and Robins et al. (2000) introduced structural nested mean models (SNMMs) and marginal structural models (MSMs), respectively. Compared to the g-formula, both of them may be used to impose parsimonious models for heterogeneous treatment effects. However, while SNMMs model how the effect of treatment at each time point is modified by the entire past covariate and treatment history, MSMs are only able to model effect modification by covariates measured at start of follow up. Furthermore, a particular implementation of SNMMs estimates the optimal treatment strategy among all possible strategies (that depend on the observed data) through a semiparametric version of dynamic programming, often referred to as A-learning in the dynamic treatment regime literature. This is in contrast to dynamic MSMs that can only be used to estimate the optimal strategy among a (smaller) prespecified class of regimes (Murphy, 2003; Robins, 2004; Henderson et al., 2010; Schulte et al., 2014; Shi et al., 2018; Qian et al., 2021). Consequently, SNMMs have gained popularity among practitioners in recent years because of the growing interest in discovering effect heterogeneity.

In the simplest case where there is only one follow-up, the SNMM is known as the structural mean model (SMM) and takes the following form:

$$g(E[Y(1) \mid A_0 = 1, L_0 = l_0]) - g(E[Y(0) \mid A_0 = 1, L_0 = l_0]) = B(l_0; \alpha), \tag{2}$$

where g is the link function, $Y(a_0), a_0 = 0, 1$ denotes the potential outcome and $B(l_0; \alpha)$ is a function known up to a finite-dimensional parameter α such that $B(l_0; 0) = 0$. A leading special case is the linear specification $B(l_0; \alpha) = \alpha^T l_0$. Under the sequential ignorability assumption (Robins, 1986), the structural model (2) and the linear specification for $B(l_0; \alpha)$ imply the following observed data model:

$$g(E[Y \mid A_0 = 1, L_0 = l_0]) - g(E[Y \mid A_0 = 0, L_0 = l_0]) = \alpha^T l_0. \tag{3}$$

Model (3) is semiparametric as it does not specify the full regression model $g(E[Y \mid A_0 = a_0, L_0 =$

$l_0]$). To enable maximum likelihood estimation, one may assume an additional baseline mean model, $E[Y \mid A_0 = 0, L_0 = l_0; \zeta]$, resulting in a generalized linear model (GLM). Alternatively, with the log or identity link, estimation of α is often based on doubly robust g-estimation methods.

An outstanding problem in the application of SNMMs to realistic setting is that when the outcome Y is binary and the link g is the log or the identity function, even for the simple point exposure case, a baseline mean model $P(Y = 1 \mid A_0 = 0, L_0 = l_0; \zeta)$ is variation dependent on, and hence can be incompatible with, the SMM. In this case, maximum likelihood estimation requires constrained optimization in a restricted parameter space, and with a new covariate value for L_0, the maximum likelihood estimator (MLE) $(\hat{\alpha}_{\text{mle}}, \hat{\zeta}_{\text{mle}})$ may still imply a fitted risk $\hat{P}(Y = 1 \mid A_0 = 1, L_0 = 0)$ to be greater than one. Additionally, the g-estimators fail to be "truly doubly robust" (Wang et al., 2021) because of the incompatible baseline model. On the other hand, when the link g is the logistic function, it is not possible to use g-estimation methods for estimating parameters in SNMMs (Robins, 2000). In particular, unlike the case with the additive or multiplicative SNMM, it is not possible to guarantee consistent estimation of logistic SNMM parameters even in randomized trials (Robins and Rotnitzky, 2004). Inference for logistic SNMMs is also considerably more complicated than that for the multiplicative or additive SNMMs; see, for example, Vansteelandt and Goetghebeur (2003); Robins and Rotnitzky (2004) and Matsouaka and Tchetgen Tchetgen (2014). Furthermore, the logistic SNMMs estimate odds ratios which are not collapsible (Rothman et al., 2008). The non-collapsibility of odds ratios limits the interpretability and generalizability of estimates from logistic SNMMs.

For the reasons mentioned above, over the past two decades SNMMs were regarded as inappropriate for inferring causal effects with binary outcomes (e.g. Robins, 2000; Daniel et al., 2013; Vansteelandt, 2010; Vansteelandt and Joffe, 2014). Richardson et al. (2017) offered a novel approach to overcoming these problems in the point exposure case, with a binary treatment and binary outcome. The key observation is that the baseline risk model included in a GLM is often not of primary interest;

instead, it is a *nuisance* model to aid estimation of the SMM parameter. To resolve the variation dependence between the *conventional* nuisance model and the SMM, they introduce a novel nuisance model that is variation independent of the SMM with the log or identity link. In conjunction with the SMM, their nuisance model gives rise to a likelihood for $P(Y = 1 \mid A_0 = a_0, L_0 = l_0)$ so that one can use unconstrained maximum likelihood for estimation. Furthermore, it permits true doubly robust g-estimation as the nuisance model is compatible with the SMM. In a recent paper, Yin et al. (2021) extended their method to accommodate a categorical exposure or, under an additional monotonicity assumption, a continuous exposure.

In this paper, we study the more challenging case of time-varying treatments. To focus on the main ideas, we primarily discuss the special case in which the time-varying treatments, time-varying confounders and outcome are all binary; note that the baseline confounders are still allowed to be continuous or discrete. This special case is prevalent in modern applications such as micro-randomized trials considered by Qian et al. (2021), where the time-varying confounder denotes the availability of a participant for the treatment. We also discuss extensions to accommodate categorical or (under additional assumptions) continuous time-varying treatments and confounders in Section S1 of the Supplementary Material.

Specifically, we show that unlike in the point exposure case, the causal parameters of binary additive SNMMs are generally variation *dependent* on each other. In comparison, the causal parameters of binary multiplicative SNMMs are variation *independent* of each other. For the latter, in parallel to the work of Richardson et al. (2017), we develop novel nuisance models that are variation independent of the multiplicative SNMMs. This is more challenging than in the point exposure problem as we need to ensure compatibility among a much larger set of models: the number of these models grows exponentially with the number of time points considered.

The rest of this article is organized as follows. In Section 2 we review the work of Richardson et al. (2017) on the point exposure case, as well as that of Robins (1994) on SNMMs. In Section

3 we present our main results on parameterizations for the binary multiplicative and additive SNMMs. We summarize our estimation approaches in Section 4, and then illustrate our approach via simulations and an application to the Mothers' Stress–Children's Morbidity Study in Sections 5 and 6, respectively. We end with a discussion in Section 7.

2. Framework and problem description

2.1 *Review of coherent models for the relative risk and risk difference*

Consider a biomedical study with binary treatment A_0, outcome Y and general baseline covariates L_0. Under the conditional ignorability condition that $A_0 \perp\!\!\!\perp Y(a_0) \mid L_0$, the SMM (2) is equivalent to the observed data model (3). With a continuous Y, a common approach to estimate the SMM parameter α is to assume a GLM,

$$g(E[Y \mid A_0, L_0]) = \zeta^T L_0 + (\alpha^T L_0)A_0 \tag{4}$$

which is equivalent to assuming model (3) and an additional nuisance model,

$$g(E[Y \mid A_0 = 0, L_0]) = \zeta^T L_0. \tag{5}$$

The GLM (4) with $g(x) = x$ and $g(x) = \log(x)$, specify, respectively, linear and log-linear models for the mean. Inference for α (and ζ) may then be based on the likelihood function $L(Y \mid A_0, L_0; \alpha, \zeta)$.

Now consider the case where Y is binary; the left hand side of (3) is known as the conditional risk difference when $g(x) = x$, and the conditional log relative risk $\log(RR(l_0))$ when $g(x) = \log(x)$. Linear or log-linear regression is often regarded inappropriate for binary Y because the nuisance model (5) is variation dependent on the model of interest (3), in the sense that the range of α depends on the specific value of ζ. For example, suppose that the baseline risk $E[Y \mid A_0 = 0, L_0 = l_0] = 0.5$, then $RD(l_0) \in [-0.5, 0.5]$ and $RR(l_0) \leqslant 2$. The variation dependence has led to a series of problems in interpretation, estimation and computation. As a result, the multiplicative

and additive SMMs have been considered inappropriate for dealing with a binary outcome (e.g. Daniel et al., 2013).

To solve this problem, Richardson et al. (2017) introduce a novel nuisance function, the l_0-specific odds product:

$$OP(l_0) = \frac{p_0(l_0)p_1(l_0)}{(1-p_0(l_0))(1-p_1(l_0))}$$

and show that it is variation independent of both the $RD(l_0)$ and $\log(RR(l_0))$; here $p_{a_0}(l_0) = P(Y = 1 \mid A_0 = a_0, L_0 = l_0)$. Furthermore, given l_0, the mappings

$$(RD(l_0), OP(l_0)) \to (p_1(l_0), p_0(l_0)) \quad \text{and} \quad (\log RR(l_0), OP(l_0)) \to (p_1(l_0), p_0(l_0)) \qquad (6)$$

are both smooth bijections from $\mathbb{D} \times \mathbb{R}^+$ to $(0,1)^2$, where $\mathbb{R}^+ = (0, \infty)$, $\mathbb{D} = (-1, 1)$ for $RD(l_0)$ and $\mathbb{D} = \mathbb{R}$ for $\log(RR(l_0))$. Figure 3 gives an illustration. One can see that each contour line of relative risk or risk difference intersects with a given contour line of the odds product at one and only one point, so that these parameters are variation independent *and* the maps in (6) are bijections. The latter feature is important as it allows for likelihood inference and risk predictions with these models.

[Figure 3 about here.]

2.2 *Structural nested mean models: Introduction*

In this paper we consider a more general biomedical study with longitudinal measurements at multiple time points $0, \ldots, K$. Let $L_k; k = 1, \ldots, K$ and $A_k; k = 0, \ldots, K$ denote the binary covariate measurement and binary treatment indicator at time k, respectively, and let Y be the outcome measured at time $K + 1$; recall that L_0 denotes general baseline covariates that can be possibly high-dimensional and are not necessarily binary. The assumption on binary A_k and L_k may be relaxed; see Section S1 in the Supplementary Material. We also let $\bar{A}_k$ and $\bar{L}_k$ denote the treatment and covariate history up to time k, that is $\bar{A}_k = (A_0, \ldots, A_k)$ and $\bar{L}_k = (L_0, \ldots, L_k)$. We presume that the covariate measurement precedes treatment at the same time point. We use

$Y(\bar{a}_k, \mathbf{0})$ to denote the outcome that would have been observed had the subject been exposed to treatment $\bar{a}_k$ until time k and treatment 0 thereafter. Implicit in this notation is the assumption of no interference between different subjects.

Following Robins (1986), we make the sequential ignorability assumption such that

$$A_k \perp\!\!\!\perp Y(\bar{a}_K) \mid \bar{L}_k, \bar{A}_{k-1} = \bar{a}_{k-1} \tag{7}$$

for all $\bar{a}_K \in \{0,1\}^K$. We also make the positivity assumption such that

$$P(A_k = a_k \mid \bar{L}_k = \bar{l}_k, \bar{A}_{k-1} = \bar{a}_{k-1}) \in (0,1), k = 0, \ldots, K \tag{8}$$

as long as $P(\bar{L}_k = \bar{l}_k, \bar{A}_{k-1} = \bar{a}_{k-1}) > 0$.

SNMMs model the conditional causal contrasts (Robins, 1994):

$$\begin{aligned} B(\bar{l}_k, \bar{a}_{k-1}; \boldsymbol{\alpha}) \equiv\ & g\{E(Y(\bar{a}_{k-1}, 1, \mathbf{0}) \mid \bar{L}_k = \bar{l}_k, \bar{A}_{k-1} = \bar{a}_{k-1}; \boldsymbol{\alpha})\} \\ & - g\{E(Y(\bar{a}_{k-1}, 0, \mathbf{0}) \mid \bar{L}_k = \bar{l}_k, \bar{A}_{k-1} = \bar{a}_{k-1}; \boldsymbol{\alpha})\} \end{aligned} \tag{9}$$

for $k = 0, \ldots, K$, where $B(\bar{l}_k, \bar{a}_{k-1}; \mathbf{0}) = 0$. The contrasts (9) are called *causal blip functions* as they describe the effect of receiving a last 'blip' of treatment at time k and then not receiving treatment thereafter (versus not receiving treatment at times $k, \ldots, K$).

SNMMs are often used for analysis of dynamic treatment regimes, where a dynamic regime is one in which a subject's treatment choice A_k depends on the intermediate responses up to that point $\bar{L}_k$ and previous treatment history $\bar{A}_{k-1}$. In fact, under sequentially ignorability (7) and the positivity assumption (8), the so-called g-null hypothesis

$$\mathcal{H}_0 : E[Y(g)] = E[Y] \quad \text{for all } g \in \mathbb{G}, \tag{10}$$

is equivalent to all the contrasts in (9) being equal to 0, where $\mathbb{G}$ denotes the set of all generalized treatment regimes consisting of all non-dynamic and dynamic treatment regimes (Robins, 1994). This statement holds regardless of the specific modeling assumptions placed on (9), since the SNMMs are guaranteed to be correctly specified under the g-null.

2.3 *Structural nested mean models: Estimation*

In practice, the SNMM parameters are often estimated using doubly robust g-estimation methods (Vansteelandt and Joffe, 2014); see Section S7.2 in the Supplementary Material for a detailed discussion. Alternatively, they may be estimated using regression-based methods, in a way similar to the GLM approach for estimating the SMM parameters. As an illustration, we discuss regression-based inference for the multiplicative SNMM with two time points and continuous outcome.

When $K = 1$, the multiplicative SNMMs model the following causal blip functions:

$$\text{(causal blips)} \quad \theta_0(l_0) \equiv \frac{E[Y(1,0) \mid L_0 = l_0]}{E[Y(0,0) \mid L_0 = l_0]}, \tag{11}$$

$$\theta_1(l_0, a_0, l_1) \equiv \frac{E[Y(a_0,1) \mid L_0 = l_0, A_0 = a_0, L_1 = l_1]}{E[Y(a_0,0) \mid L_0 = l_0, A_0 = a_0, L_1 = l_1]} = \frac{E[Y \mid l_0, a_0, l_1, 1]}{E[Y \mid l_0, a_0, l_1, 0]}. \tag{12}$$

To allow for regression modeling on $E[Y \mid l_0, a_0, l_1, a_1]$, one may specify additional nuisance models on the following functions (Robins, 1997, Appendix 2, p.36):

$$(L_1 \text{ blip}) \quad \widetilde{\phi}(l_0, a_0, l_1) \equiv \frac{E[Y(a_0,0) \mid L_0 = l_0, A_0 = a_0, L_1 = l_1]}{E[Y(a_0,0) \mid L_0 = l_0, A_0 = a_0, L_1 = 0]} = \frac{E[Y \mid l_0, a_0, l_1, 0]}{E[Y \mid l_0, a_0, 0, 0]}; \tag{13}$$

$$\widetilde{\eta}(l_0, a_0) \equiv f(L_1 \mid L_0 = l_0, A_0 = a_0); \tag{14}$$

$$\widetilde{\psi}(l_0) \equiv E(Y \mid L_0 = l_0, A_0 = 0, L_1 = 0, A_1 = 0). \tag{15}$$

Note that the "L_1 blip" is not causal. Furthermore, with these nuisance models, the mean potential outcome $E[Y(a_0, a_1)]$ may be evaluated via the g-formula (1).

3. Parameterizations of binary SNMMs

We now consider the case with a binary outcome Y. In this case, regression models on $E[Y \mid l_0, a_0, l_1, a_1]$ and $f(L_1 \mid l_0, a_0)$ give rise to likelihood functions. Hence, in principle, one may use likelihood-based inference for inferring the SNMM parameters. However, similar to the case of SMMs discussed in Section 2.1, with a binary Y, the models for (11) – (15) are variation dependent on each other, leading to undesirable consequences for interpretation, estimation and computation. These problems may be avoided if (I) the SNMMs are variation independent of each other; (II) the

nuisance models are variation independent of the SNMMs; (III) there exists a bijection between the regression models on $E[Y \mid l_0, a_0, l_1, a_1]$, $f(L_1 \mid l_0, a_0)$ and the combination of SNMMs and nuisance models. In Section 3.1, we show that (I) is true for multiplicative SNMMs but not for additive SNMMs. As a result, in general, estimators for the additive SNMM parameters may not be obtained via unconstrained maximum likelihood estimation. On the other hand, to construct an unconstrained MLE for the multiplicative SNMM parameters, in Section 3.2 and 3.3, we propose novel nuisance functions that satisfy criteria (II) and (III).

3.1 *Variation independence of SNMM parameters*

PROPOSITION 1: If $K \geqslant 1$, then the additive SNMMs are *variation dependent* on each other, while the multiplicative SNMMs are variation independent of each other.

Proposition 1 may be surprising at first sight. To provide heuristics, we illustrate the result in the case $K = 1$. A formal proof will become obvious later given Theorem 2.

When $K = 1$, the additive SNMMs model a sequence of contrasts including

$$E[Y(1,0) - Y(0,0) \mid L_0 = l_0] \in (-1, 1), \tag{16}$$

$$E[Y(1,1) - Y(1,0) \mid L_0 = l_0, A_0 = 1, L_1 = l_1] \in (-1, 1). \tag{17}$$

Marginalizing (17) over the distribution of L_1 conditional on $L_0 = l_0, A_0 = 1$ and using the sequential ignorability assumption (7), we get

$$E[Y(1,1) - Y(1,0) \mid L_0 = l_0] \in (-1, 1). \tag{18}$$

If (16) were variation independent of (17) (and hence (18)), then the range of the sum of (16) and (18) would be $(-2, 2)$. This contradicts the fact that the range of $E[Y(1,1) - Y(0,0) \mid L_0 = l_0]$ is $(-1, 1)$.

The reasoning above does not constitute a contradiction for the multiplicative SNMM as it specifies a sequence of differences of the form $\log\{E[Y(\bar{a}_k, 0) \mid \bar{L}_k = \bar{l}_k, \bar{A}_k = \bar{a}_k]\} - \log\{E[Y(\bar{a}_{k-1}, 0) \mid \bar{L}_k = \bar{l}_k, \bar{A}_k = \bar{a}_k]\} \in \mathbb{R}$. In simple terms, the multiplicative SNMMs are variation independent as

$\mathbb{R}+\mathbb{R}=\mathbb{R}$, whereas the additive SNMMs are variation *dependent* as $(-1,1)+(-1,1) \not\subset (-1,1)$; here interval additions are defined as $(x_1,x_2)+(y_1,y_2)=\{x+y \mid x\in(x_1,x_2), y\in(y_1,y_2)\}=(x_1+y_1,x_2+y_2)$.

We can also explain Proposition 1 with the graphs in Figure 4. As shown in Figure 4 (a) and (c), with additive SNMMs, the second stage blips $E[Y(a_0,1)-Y(a_0,0) \mid L_0=l_0, A_0=a_0, L_1=l_1], l_1=0,1$ may impose constraints on the second stage baseline quantities $E[Y(a_0,0) \mid L_0=l_0, A_0=a_0, L_1=l_1]$. Marginalizing over the distribution of L_1 given L_0, A_0, these constraints may then imply constraints on $E[Y(a_0,0) \mid L_0=l_0, A_0=a_0]$, which, by the sequential ignorability assumption, equals $E[Y(a_0,0) \mid L_0=l_0]$. The constraints on $E[Y(a_0,0) \mid L_0=l_0]$, $a_0=0,1$ are shown in the red and blue rectangles in Figure 4 (b), respectively. The intersection of these rectangles, i.e. the dotted region in Figure 4 (b), defines the constraints on $(E[Y(0,0) \mid L_0=l_0], E[Y(1,0) \mid L_0=l_0])$ implied by the second stage blips. One may see that some contour lines of $E[Y(1,0)-Y(0,0) \mid L_0=l_0]$, such as the gray lines in Figure 4 (b), have no intersection with the dotted feasible region for $(E[Y(0,0) \mid L_0=l_0], E[Y(1,0) \mid L_0=l_0])$. This shows that the second stage blips may imply constraints on the possible values of the first stage blip $E[Y(1,0)-Y(0,0) \mid L_0=l_0]$, so that they are variation dependent of each other.

We can apply the same reasoning to the multiplicative SNMMs, as illustrated in Figure 5. Given any values for the second stage blips, the feasible region for $(E[Y(0,0) \mid L_0=l_0], E[Y(1,0) \mid L_0=l_0])$ always includes the origin. Hence, unlike the case with additive SNMMs, this feasible region will always intersect with any contour lines for the first stage blip $E[Y(0,1) \mid L_0=l_0]/E[Y(0,0) \mid L_0=l_0]$.

[Figure 4 about here.]

[Figure 5 about here.]

3.2 *Parameterization for the multiplicative SNMM: The two time points case*

We now discuss the choice of nuisance models for the binary multiplicative SNMM with two time points. We first note that given $\widetilde{\eta}(l_0, a_0)$, the SNMM parameters $\theta_0(l_0), \theta_1(l_0, a_0, l_1)$ and L_1 blip function $\widetilde{\phi}(l_0, a_0, l_1)$ imply the seven conditional relative risks $E(Y \mid l_0, a_0, l_1, a_1)/E(Y \mid l_0, 0, 0, 0), (a_0, l_1, a_1) \in \{0,1\}^3 \setminus \{(0,0,0)\}$; see eqn. (S3) in the Supplementary Material. As we discussed in Section 2.1, these conditional relative risks are variation dependent on the baseline risk $\widetilde{\psi}(l_0)$.

To solve this problem, motivated by the l_0-specific odds product in Richardson et al. (2017), we propose to replace the baseline risk function $\widetilde{\psi}(l_0)$ in (15) with the l_0-specific generalized odds product:

$$\text{gop}(l_0) \equiv \frac{\prod_{a_0=0,1} \prod_{l_1=0,1} \prod_{a_1=0,1} E[Y \mid l_0, a_0, l_1, a_1]}{\prod_{a_0=0,1} \prod_{l_1=0,1} \prod_{a_1=0,1} (1 - E[Y \mid l_0, a_0, l_1, a_1])}.$$

Note that similar to the l_0-specific odds product, the l_0-specific generalized odds product is only a function of l_0 and not (a_0, l_1, a_1). Theorem 1 shows that replacing $\widetilde{\psi}(l_0)$ with $\text{gop}(l_0)$ gives rise to a variation independent parameterization.

THEOREM 1: *Suppose that the sequential ignorability assumption* (7) *holds. Let* $\mathcal{M}$ *denote the model specified by the SNMMs* (11) *and* (12)*, a model on gop*(l_0)*, and models on the following nuisance parameters:*

$$\begin{aligned} &(L_1 \text{ blip}) \quad \phi(l_0, a_0) \equiv \widetilde{\phi}(l_0, a_0, 1), a_0 = 0, 1, \\ &\qquad\qquad \eta(l_0, a_0) \equiv E[L_1 \mid L_0 = l_0, A_0 = a_0], a_0 = 0, 1. \end{aligned} \tag{19}$$

Then for each l_0*, the map given by*

$$\begin{aligned} &(\theta_0(l_0), \theta_1(l_0, 1, 1), \theta_1(l_0, 1, 0), \theta_1(l_0, 0, 1), \theta_1(l_0, 0, 0), \phi(l_0, 0), \phi(l_0, 1), \text{gop}(l_0), \eta(l_0, 0), \eta(l_0, 1)) \to \\ &\qquad (P(Y = 1 \mid l_0, a_0, l_1, a_1), a_0, l_1, a_1 \in \{0,1\}; \eta(l_0, 0), \eta(l_0, 1)) \end{aligned} \tag{20}$$

is a bijection from $(\mathbb{R}^+)^8 \times (0,1)^2$ *to* $(0,1)^{10}$. *Furthermore, models defining* $\mathcal{M}$ *are variation independent of each other.*

Theorem 1 may be generalized to accommodate categorical or (under additional monotonicity conditions) continuous time-varying treatments A_0, A_1 and confounder L_1; see Section S1 in the Supplementary Material for details.

3.3 *Parameterization for the multiplicative SNMM: The general case*

To describe parameterizations for the binary multiplicative SNMM in the general case, we first introduce some notation:

$$\vec{0}_k \equiv \overbrace{(0,\ldots,0)}^{k\ 0s};$$

$$\begin{aligned}
\text{for } k=0,\ldots,K: \quad & E[Y(\bar{a}_k,\mathbf{0})] \equiv E[Y(\bar{a}_k,\vec{0}_{K-k})];\\
\text{for } k=0,\ldots,K: \quad & \theta_k(\bar{a}_{k-1},\bar{l}_k) \equiv \frac{E[Y(\bar{a}_{k-1},1,\mathbf{0}) \mid \bar{A}_{k-1}=\bar{a}_{k-1},\bar{L}_k=\bar{l}_k]}{E[Y(\bar{a}_{k-1},0,\mathbf{0}) \mid \bar{A}_{k-1}=\bar{a}_{k-1},\bar{L}_k=\bar{l}_k]};\\
\text{for } k=0,\ldots,K-1: \quad & \phi(\bar{a}_k,\bar{l}_k) \equiv \frac{E[Y(\bar{a}_k,\mathbf{0}) \mid \bar{A}_k=\bar{a}_k,\bar{L}_k=\bar{l}_k,L_{k+1}=1]}{E[Y(\bar{a}_k,\mathbf{0}) \mid \bar{A}_k=\bar{a}_k,\bar{L}_k=\bar{l}_k,L_{k+1}=0]};\\
\text{for } k=0,\ldots,K-1: \quad & \eta(\bar{a}_k,\bar{l}_k) \equiv E[L_{k+1} \mid \bar{A}_k=\bar{a}_k,\bar{L}_k=\bar{l}_k].
\end{aligned}$$

We also let $\bar{a}_{-1}=\emptyset$ so that when $k=0$,

$$\theta_k(\bar{a}_{-1},\bar{l}_0) \equiv \frac{E[Y(1,\mathbf{0}) \mid L_0=l_0]}{E[Y(0,\mathbf{0}) \mid L_0=l_0]} \quad \text{and } \phi(\bar{a}_0,\bar{l}_0) \equiv \frac{E[Y(a_0,\mathbf{0}) \mid A_0=a_0,L_0=l_0,L_1=1]}{E[Y(a_0,\mathbf{0}) \mid A_0=a_0,L_0=l_0,L_1=0]}.$$

The following Theorem 2 gives the general form of our nuisance parameters for binary multiplicative SNMMs.

THEOREM 2: *Suppose that the sequential ignorability assumption* (7) *holds. Let* $\mathcal{M}$ *denote the model specified by the SNMMs on*

$$\textit{(Stage-k causal blip)} \quad \boldsymbol{\theta} = (\theta_k(\bar{a}_{k-1},\bar{l}_k) : k=0,\ldots,K)$$

and models on the following nuisance parameters

$$\begin{aligned}
(L_{k+1}\ \textit{blip}) \quad \boldsymbol{\phi} &= (\phi(\bar{a}_k, \bar{l}_k) : k = 0, \ldots, K-1); \\
\boldsymbol{\eta} &= (\eta(\bar{a}_k, \bar{l}_k) : k = 0, \ldots, K-1); \\
GOP(l_0) &\equiv \frac{\prod\limits_{\bar{a}_K, l_1, \ldots, l_K} E[Y \mid \bar{A}_K = \bar{a}_K, \bar{L}_K = \bar{l}_K]}{\prod\limits_{\bar{a}_K, l_1, \ldots, l_K} (1 - E[Y \mid \bar{A}_K = \bar{a}_K, \bar{L}_K = \bar{l}_K])}.
\end{aligned}$$

Then for each l_0, the map given by

$$(\boldsymbol{\theta}, \boldsymbol{\phi}, GOP, \boldsymbol{\eta}) \rightarrow (E[Y \mid \bar{A}_K = \bar{a}_K, \bar{L}_K = \bar{l}_K], \boldsymbol{\eta}) \tag{21}$$

is a bijection from $(\mathbb{R}^+)^{d_1} \times (0,1)^{d_2}$ to $(0,1)^d$, where $d_1 = 2^{2K+1}, d_2 = \sum\limits_{k=0}^{K-1} 2^{2k+1}, d = d_1 + d_2$. Furthermore, models in $\mathcal{M}$ are variation independent of each other.

4. Estimation

4.1 *Maximum likelihood estimation: The two time points case*

We first discuss maximum likelihood estimation for a multiplicative SNMM with two time points. Suppose models in $\mathcal{M}$ are specified up to a finite dimensional parameter, then the parameters may be estimated directly via unconstrained maximum likelihood based on the bijection (20). For example, in the simulations, we assume that

$$\theta_0(l_0) = \exp(\alpha_j^T l_0), j = 0; \tag{22}$$

$$\theta_1(l_0, j) = \exp(\alpha_j^T l_0), j = (1,1), (1,0), (0,1), (0,0); \tag{23}$$

$$\phi(l_0, j) = \exp(\beta_j^T l_0), j = 0, 1; \tag{24}$$

$$\text{gop}(l_0) = \exp(\delta^T l_0), \tag{25}$$

$$\eta(l_0, j) = \text{expit}(\gamma_j^T l_0), j = 0, 1; \tag{26}$$

For every possible value of $\alpha_j, \beta_j, \delta, \gamma_j$, we may use equations (22) – (26) and the following Algorithm 1 to compute $E(Y \mid l_0, a_0, l_1, a_1)$. The likelihood is then calculated as $\prod\limits_{i=1}^{n} L_i^Y \times L_i^L$, where $L_i^Y = E(Y \mid L_{0i}, A_{0i}, L_{1i}, A_{1i})^{Y_i} \{1 - E(Y \mid L_{0i}, A_{0i}, L_{1i}, A_{1i})\}^{1-Y_i}$ and $L_i^L = E(L_1 \mid$

$L_{0i}, A_{0i})^{L_{1i}} \{1 - E(L_1 \mid L_{0i}, A_{0i})\}^{1-L_{1i}}$. The unconstrained maximum likelihood estimate of $(\alpha_j, \beta_j, \delta, \gamma_j)$ can then be obtained by directly maximizing the log likelihood.

Algorithm 1 Compute $E[Y \mid l_0, a_0, l_1, a_1]$ from $(\theta_0(l_0), \theta_1(l_0, a_0, l_1), \phi(l_0, a_0), gop(l_0), \eta(l_0, a_0))$

(1) Compute $\dfrac{E[Y \mid l_0, a_0, l_1, 1]}{E[Y \mid l_0, a_0, l_1, 0]}$, $\dfrac{E[Y \mid l_0, a_0, 1, 0]}{E[Y \mid l_0, a_0, 0, 0]}$ and $\dfrac{E[Y \mid l_0, 1, 0, 0]}{E[Y \mid l_0, 0, 0, 0]}$ via (11), (12), (19) and

$$\begin{aligned}\theta_0(l_0) &= \frac{\eta(l_0,1)E[Y \mid l_0,1,1,0] + (1-\eta(l_0,1))E[Y \mid l_0,1,0,0]}{\eta(l_0,0)E[Y \mid l_0,0,1,0] + (1-\eta(l_0,0))E[Y \mid l_0,0,0,0]} \\ &= \frac{\eta(l_0,1)\phi(l_0,1) + 1 - \eta(l_0,1)}{\eta(l_0,0)\phi(l_0,0) + 1 - \eta(l_0,0)} \times \frac{E[Y \mid l_0,1,0,0]}{E[Y \mid l_0,0,0,0]}.\end{aligned}$$

(2) Compute $r_{a_0,l_1,a_1}(l_0) \equiv \dfrac{E[Y \mid l_0, a_0, l_1, a_1]}{E[Y \mid l_0, 0, 0, 0]}$ sequentially by

$$\begin{aligned}\frac{E[Y \mid l_0,a_0,0,0]}{E[Y \mid l_0,0,0,0]} &= \frac{E[Y \mid l_0,a_0,0,0]}{E[Y \mid l_0,0,0,0]} \quad \text{obtained from step (1);} \\ \frac{E[Y \mid l_0,a_0,l_1,0]}{E[Y \mid l_0,0,0,0]} &= \frac{E[Y \mid l_0,a_0,1,0]}{E[Y \mid l_0,a_0,0,0]} \times \frac{E[Y \mid l_0,a_0,0,0]}{E[Y \mid l_0,0,0,0]}; \\ \frac{E[Y \mid l_0,a_0,l_1,a_1]}{E[Y \mid l_0,0,0,0]} &= \frac{E[Y \mid l_0,a_0,l_1,a_1]}{E[Y \mid l_0,a_0,l_1,0]} \times \frac{E[Y \mid l_0,a_0,l_1,0]}{E[Y \mid l_0,0,0,0]}.\end{aligned}$$

(3) Compute $r_{\max}(l_0) = \max_{a_0,l_1,a_1} r_{a_0,l_1,a_1}(l_0)$.

(4) Compute $k_{a_0,l_1,a_1}(l_0) \equiv \dfrac{r_{a_0,l_1,a_1}(l_0)}{r_{\max}(l_0)}$.

(5) Let $p_{l_0,a_0,l_1,a_1} = E[Y \mid l_0, a_0, l_1, a_1]$. Suppressing dependence on l_0, for $x \in (0,1)$, let

$$g(x) = \sum_{i=(a_0,l_1,a_1)} \log(k_i) + 8\log(x) - \sum_{i=(a_0,l_1,a_1)} \log(1 - k_i x) - \log(gop).$$

Find the unique root of $g(x)$ in the interval $(0,1)$. Set $p_{\max}(l_0)$ to be this root.

(6) Compute $E[Y \mid l_0, a_0, l_1, a_1] = k_{a_0,l_1,a_1}(l_0) \times p_{\max}(l_0)$.

Alternatively, a two-step procedure may be employed, in which one first estimates $\gamma_j, j = 0, 1$, by maximizing the likelihood associated with $P(L_1 = 1 \mid A_0, L_0)$, that is, $\prod_{i=1}^{n} L_i^L$. The value of $\gamma_j, j = 0, 1$ are then taken as fixed before proceeding to find the partial maximum likelihood estimate of $\alpha_j, \beta_j, \delta$ that maximizes $\prod_{i=1}^{n} L_i^Y$. Inference for both the unconstrained and two-step maximum likelihood estimates can then be performed based on the non-parametric bootstrap.

4.2 *Maximum likelihood estimation: The general case*

In the general case with $K+1$ time points, in principle, the unconstrained and two-step maximum likelihood estimate may be obtained in a similar way. The corresponding algorithm to compute the mapping (21) is given in Algorithm S1 in the Supplementary Material.

However, note that in general, the dimension of model parameters $(\boldsymbol{\theta}, \boldsymbol{\phi}, GOP, \boldsymbol{\eta})$ grows exponentially with K. To avoid possible identification problems with large numbers of follow-ups and moderate sample sizes, in practice one may make further dimension reducing assumptions on $\mathcal{M}$. In the data application, we make the Markov assumption that $\boldsymbol{\theta}, \boldsymbol{\phi}$ and $\boldsymbol{\eta}$ depend on the past history only through the most recent L_k, A_k and (A_k, L_k), respectively, and assume that such dependencies are homogeneous over time:

$$\theta_k(\bar{a}_{k-1}, \bar{l}_k) = \theta(l_k), \quad \phi(\bar{a}_k, \bar{l}_k) = \phi(a_k), \quad \eta(\bar{a}_k, \bar{l}_k) = \eta(a_k, l_k). \tag{27}$$

Similar assumptions have also been invoked in recent work by Qian et al. (2021). The Markov assumption may be extended to allow dependence on the previous two time points; see Section S6 in the Supplementary Material for details.

Furthermore, since the dimension of $E[Y \mid \bar{A}_K, \bar{L}_K]$ grows exponentially with K, Algorithm S1, and specifically Steps 2 – 5 may be computationally prohibitive even when K is small to moderate. To resolve this problem, instead of computing $r_{\bar{a}_K, \bar{l}_K}$ for each $(\bar{a}_K, \bar{l}_K)$, we develop a dynamic programming method to compute the exact value of $r_{\max}(l_0)$; the details are provided in Section S5 of the Supplementary Material. Dynamic programming is applicable here due to our Markov assumption that $\boldsymbol{\theta}, \boldsymbol{\phi}$ and $\boldsymbol{\eta}$ depend on the past history only through the most recent L_k or A_k. The dynamic programming method has also been used before in the optimal structural nested models (Robins, 2004, a.k.a. A-learning) literature. Furthermore, in Step 5, we approximate $g(x)$ by

$$h_m(x) = \frac{d_1}{m} \sum_{i=1}^{m} \log(k_i) + d_1 \log(x) - \frac{d_1}{m} \sum_{i=1}^{m} \log(1 - k_i x) - \log(GOP),$$

where $i = 1, \ldots, m$ are random samples drawn from a uniform distribution on the set $\{0,1\}^{d_1}$. To choose m in practice, one may start with a small number, say $m = 100$, and then increase m until $h_m(x)$ is stable up to a threshold specified a priori. With the dynamic programming method and Monte Carlo approximation, the computational cost is reduced from $\mathcal{O}(\exp(K))$ to $\mathcal{O}(K)$.

4.3 *Doubly robust estimation*

The proposed parameterization in Theorems 1 and 2 may be used to construct truly doubly robust estimators that are asymptotically linear if either the nuisance models $\boldsymbol{\phi}(\beta), GOP(\delta), \boldsymbol{\eta}(\gamma)$ or the propensity score models $P(A_k = 1 \mid \bar{A}_{k-1}, \bar{L}_k; \epsilon)$ are correctly specified. These estimators are called "truly" doubly robust because, as shown in Theorem 1, the nuisance models $\boldsymbol{\phi}(\beta), GOP(\delta), \boldsymbol{\eta}(\gamma)$ are compatible with the causal models $\boldsymbol{\theta}(\alpha)$. To keep the exposition simple, we only discuss the case with $K = 1$ here; the results can be extended to the general case, as we illustrate in Section S7.2 in the Supplementary Material.

Specifically, let

$$\begin{aligned} U_1(\alpha) =& Y\theta_1(L_0, A_0, L_1; \alpha)^{-A_1}; \\ U_0(\alpha) =& Y\theta_1(L_0, A_0, L_1; \alpha)^{-A_1}\theta_0(L_0; \alpha)^{-A_0}. \end{aligned}$$

Also let $\hat{\alpha}, \hat{\beta}, \hat{\delta}, \hat{\gamma}$ and $\hat{\epsilon}$ be preliminary estimates of $\alpha, \beta, \delta, \gamma$ and ϵ obtained via MLE or 2-step MLE, and let $\hat{\alpha}_{dr}$ solve the following estimating equation (Robins, 1994; Vansteelandt and Joffe, 2014):

$$\begin{aligned} &\mathbb{P}_n \Big(\Big[d_0(L_0, A_0) - \hat{E}\left\{ d_0(L_0, A_0) \mid L_0 \right\} \Big] \times \Big[U_0(\alpha) - \hat{E}\left\{ U_0(\alpha) \mid L_0 \right\} \Big] + \\ &\Big[d_1(L_0, A_0, L_1, A_1) - \hat{E}\left\{ d_1(L_0, A_0, L_1, A_1) \mid L_0, A_0, L_1 \right\} \Big] \times \Big[U_1(\alpha) - \hat{E}\left\{ U_1(\alpha) \mid L_0, A_0, L_1 \right\} \Big] \Big) = 0, \end{aligned}$$

where $\mathbb{P}_n$ denotes the empirical mean operator: $\mathbb{P}_n O = \sum_{i=1}^{n} O_i/n$, $d_0(L_0, A_0)$ and $d_1(L_0, A_0, L_1, A_1)$ are measurable functions of the same dimension as α, and

$$\hat{E}\left\{ U_1(\alpha) \mid L_0, A_0, L_1 \right\} = \hat{E}\{ Y(A_0, 0) \mid L_0, A_0, L_1 \} = E(Y \mid L_0, A_0, L_1, A_1 = 0; \hat{\alpha}, \hat{\beta}, \hat{\delta}, \hat{\gamma});$$

$$\hat{E}\left\{U_0(\alpha) \mid L_0\right\} = \hat{E}\{Y(0,0) \mid L_0\} = \eta(0;\hat{\gamma})E(Y \mid L_0, A_0 = 0, L_1 = 1, A_1 = 0; \hat{\alpha}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) + \{1 - \eta(0;\hat{\gamma})\}E(Y \mid L_0, A_0 = 0, L_1 = 0, A_1 = 0; \hat{\alpha}, \hat{\beta}, \hat{\delta}, \hat{\gamma}). \tag{28}$$

Note that for $k = 0, 1$, the terms $\hat{E}\left\{d_k(\bar{L}_k, \bar{A}_k) \mid \bar{L}_k, \bar{A}_{k-1}\right\}$ depend on the propensity score estimates $P(A_k = 1 \mid \bar{L}_k, \bar{A}_{k-1}, \hat{\epsilon})$.

Robins (1994) showed that under correct models for $\boldsymbol{\theta}(\alpha)$ and additional regularity conditions, $\hat{\alpha}_{dr}$ is consistent and asymptotically normally distributed provided that either the models for $E\left\{U_k(\alpha) \mid \bar{L}_k, \bar{A}_{k-1}\right\}$ or $P(A_k = 1 \mid \bar{L}_k, \bar{A}_{k-1})$ are correctly specified. The optimal choice of $d_0(L_0, A_0)$ and $d_1(L_0, A_0, L_1, A_1)$ are given in Section S7.1 of the Supplementary Material.

5. Simulation studies

We now evaluate the finite sample performance of our estimators with synthetic data. In our simulations, we consider two choices for the baseline covariates L_0: "binary L_0" that includes an intercept and a binary random variable generated from a Bernoulli distribution with mean $1/2$, or "continuous L_0" that includes an intercept and a random draw from the uniform distribution on $[-2, 2]$. Conditional on L_0, the treatments A_0, A_1 and intermediate covariate L_1 were generated from (26) and the following models:

$$P(A_0 = 1 \mid L_0) = \text{expit}(\epsilon_1^T L_0); \tag{29}$$

$$P(A_1 = 1 \mid L_1, A_0, L_0) = \text{expit}(\epsilon_2^T L_0 + \epsilon_3 A_0 + \epsilon_4 L_1), \tag{30}$$

where $\epsilon_1 = \epsilon_2 = (0.1, -0.5)^T, \epsilon_3 = 0.1, \epsilon_4 = -0.5, \gamma_0 = \gamma_1 = (-0.5, 0.1)^T$. The outcome Y was generated indirectly through models (23) – (25), where $\alpha_j = (0, 0.7)^T$ for $j = 0, (0,0), (0,1), (1,0), (1,1)$; $\beta_0 = \beta_1 = (-0.5, 0.1)^T; \delta = (-0.5, 1)^T$.

We compare three estimation methods: 1) MLE as described in Section 4.1, in which we use the `R` function `optim` to maximize the likelihood, and `uniroot` to find the solution in Step (5) of

Algorithm 1; 2) 2-step MLE as described in Section 4.1; 3) DR estimation as described in Section 4.3, in which we use the 2-step MLE estimates as the preliminary estimates for $\alpha, \beta, \delta, \gamma$ and the warm starting value for α. The weighting functions are chosen as the optimal weighting function detailed in Section S7.1 of the Supplementary Material. The propensity score model parameters $\epsilon_j, j = 1, \ldots, 4$ are estimated using logistic regressions. As a summary measure, we also report estimates of the causal contrast $E[Y(1,1)]/E[Y(0,0)]$ using the g-formula (1), where $E(Y \mid l_0, a_0, l_1, a_1)$ and $f(l_1 \mid l_0, a_1)$ were estimated using the models on $\theta_0(l_0), \theta_1(l_0, a_0, l_1), \phi(l_0, a_0)$, gop$(l_0), \eta(l_0, a_0)$. The true value for $E[Y(1,1)]/E[Y(0,0)]$ is evaluated via a simulation sample of size 100,000. Unless otherwise specified, the simulation results are based on 500 Monte-Carlo runs of $n = 1000$ units.

Table 1 summarizes the simulation results for binary L_0. Results with continuous L_0 are deferred to Table S1 in the Supplementary Material. All methods yield estimators with small biases relative to their standard errors, confirming consistency of the proposed estimators. The estimates of the 2-step MLE are very close to those of the MLE, which suggests that the conditional distribution of the outcome Y, i.e. $P(Y = y \mid A_1, L_1, A_0, L_0)$ contains little information on γ_0 and γ_1 relative to $P(L_1 = l_1 \mid A_0, L_0)$. As expected, the variance of the doubly robust estimator is no smaller than that of the maximum likelihood estimators.

[Table 1 about here.]

We also compare these three methods in terms of their computation time in Table 2. All the computation was done on a Lenovo SD650 NeXtScale server using an Intel 8268 "Cascade Lake" processor and 192GB RAM. As expected, the computation time with continuous L_0 is longer as there are many more different combinations of covariate values in the sample.

[Table 2 about here.]

A reviewer suggested that with the doubly robust g-estimation procedure, when the propensity score models are correctly specified, our proposed parametrization may lead to a more efficient

estimator than a non-compatible one, even if the nuisance models are misspecified. Motivated by this, we consider an additional simulation setting with a (spurious) baseline covariate vector $\widetilde{L}_0$ that includes an intercept and an independent random draw from $Bern(1/2)$. We consider two mis-specifications for the baseline models: (1) our proposed DR estimator, with the gop model misspecified as a function of $\widetilde{L}_0$: gop$(L_0) = exp(\delta^T \widetilde{L}_0)$; note that in our approach, the baseline function $E[Y \mid L_0, A_0, L_1, A_1]$ is estimated through the gop model, so it is also misspecified; (2) the DR estimator with logistic baseline models, where the terms $E(Y \mid L_0, A_0, L_1, A_1 = 0)$ in (28) are estimated using a logistic regression that assumes logit$E[Y \mid L_0, A_0, L_1, A_1] = b_1\widetilde{L}_0 + b_1 A_0 + b_2 L_1 + b_3 A_1 L_1$. Table 3 summarizes the results. The estimates from the DR estimator with a logistic baseline model have larger variability than the proposed DR estimator, especially those for $\theta_1(l_0, 0, 1)$. This provides some initial evidence for the efficiency benefits of our proposed DR estimator, in the case that the baseline models are misspecified.

[Table 3 about here.]

In Section S8.2 of the Supplementary Material, we present an alternative simulation study that compares our proposed two-step MLE and doubly robust estimators, versus the g-computation method under the g-null hypothesis that $\theta_0(l_0) = \theta_1(l_0, a_0, l_1) = 1$ for all l_0 and $(a_0, l_1) = (1, 1), (1, 0), (0, 1), (0, 0)$.

6. Application to the Mothers' Stress–Children's Morbidity study

To illustrate the proposed methods, we reanalyze data from the Mothers' Stress-Children's Morbidity (MSCM) study (Zeger and Liang, 1986), which consist of observations on 167 mothers with infants aged between 18 months and 5 years. Daily observations were taken on mothers' stress level and whether or not their child was ill. The total length of follow-up is 30 days. The data that support the findings of this study are available from the corresponding author upon request. Similar to Robins et al. (1999), we are interested in whether or not maternal stress has an influence on

child illness. Maternal stress may be considered as a treatment variable because it is a natural target for an intervention through support programs such as by providing midwives. Following Zeger and Liang (1986), we use the first 9 days of records to illustrate use of the SNMM so that $K = 7$. The treatment variables are maternal stress indicators in the first 8 days, denoted as $A_0, \ldots, A_7$; the outcome of interest is whether or not the child is ill at the 9th day; the time-varying confounders are child illness in the first 8 days: $L_0, \ldots, L_7$, and the time-independent baseline confounders include household size, employment, marital status and child's race. To distinguish the time-independent confounders from the time-varying confounder measured at baseline, with slight abuse of notation, we use X to denote the former, and L_0 for the latter. Note that the outcomes, time-varying confounders and predictor are all binary. In the data set made available to us, the time-independent confounders are also binary. There are 147 mother-child pairs with complete observations on all these variables and for illustrative purposes, we restrict our analysis to these pairs. Figure 6 shows the observations in the first 9 days. The correlation between maternal stress and children's illness, however, may be due to confounding by children's illness at earlier time points. Our formal causal analysis assumes that all confounders are measured, and that our parametric model specifications are correct.

[Figure 6 about here.]

We assume the following causal blip models:

$$\log \frac{E[Y(\bar{a}_{k-1}, 1, \mathbf{0}) \mid \bar{A}_k = \bar{a}_k, \bar{L}_k = \bar{l}_k, X]}{E[Y(\bar{a}_{k-1}, 0, \mathbf{0}) \mid \bar{A}_k = \bar{a}_k, \bar{L}_k = \bar{l}_k, X]} = \alpha_0(1 - l_k) + \alpha_1 l_k + \alpha_X^T X, k = 0, \ldots, K$$

and the following nuisance models:

$$\log \frac{E[Y(\mathbf{0}) \mid L_0 = 1, X]}{E[Y(\mathbf{0}) \mid L_0 = 0, X]} = \beta_{L_0} + \beta_X^T X,$$

$$\log \frac{E[Y(\bar{A}_k, \mathbf{0}) \mid \bar{A}_k, \bar{L}_k, L_{k+1} = 1, X]}{E[Y(\bar{A}_k, \mathbf{0}) \mid \bar{A}_k, \bar{L}_k, L_{k+1} = 0, X]} = \beta_0(1 - A_k) + \beta_1 A_k + \beta_X^T X, \quad k = 0, \ldots, K-1;$$

$$GOP(L_0, X) = \delta_X^T X;$$

$$\text{logit}(E[L_{k+1} \mid \bar{A}_k, \bar{L}_k, X]) = \gamma_{00}1(A_k = L_k = 0) + \gamma_{01}1(A_k = 0, L_k = 1) +$$
$$\gamma_{10}1(A_k = 1, L_k = 0) + \gamma_{11}1(A_k = L_k = 1) + \gamma_X X,\ k = 0, \ldots, K-1;$$
$$\text{logit}P(A_k = 1 \mid \bar{L}_k, \bar{A}_{k-1}, X) = \epsilon_0 + \epsilon_1 L_k + \epsilon_X X,\ k = 1, \ldots, K.$$

We apply three different estimation methods: 1) 2-step MLE as described in Section 4.1; (2) DR estimation as described in Section 4.3, in which the weighting function is chosen as $d_m(\bar{L}_m, \bar{A}_m) = A_m(1, L_m, X)^T$; (3) DR estimation with a logistic baseline model:

$$\text{logit}(E[Y \mid \bar{A}_{K-1} = \bar{a}_{k-1}, A_K = 0, \bar{L}_K = \bar{l}_K, X]) = \zeta_{00}1(a_{K-1} = l_K = 0) + \zeta_{01}1(a_{K-1} = 0, l_K = 1) +$$
$$\zeta_{10}1(a_{K-1} = 1, l_K = 0) + \zeta_{11}1(a_{K-1} = l_K = 1) + \zeta_X X.$$

Inference is based on 100 non-parametric bootstrap samples. Table 4 presents the analysis results.

We first tested the g-null mean hypothesis (10). Note that as Y is binary, the g-null mean hypothesis coincides with the g-null hypothesis of Robins (1986, §6). A valid level-0.05 test may be obtained by testing $(\alpha_0, \alpha_1, \alpha_2^T) = 0$. The p-values from the three methods all suggest that we have failed to reject the g-null (mean) hypothesis at the 0.05 level. Note that, as discussed by Robins et al. (1999, §6), the standard generalized estimating equation approach of Zeger and Liang (1986) cannot be used to test the g-null hypothesis in the presence of time-dependent confounding by L_k.

We then compare estimates of the causal parameters $(\alpha_0, \alpha_1, \alpha_X)$ from the three methods. The confidence interval of the proposed DR estimator is shorter than the DR estimator with a logistic baseline model. This is consistent with the findings in Table 3, suggesting that our compatible parameterization may lead to efficiency gain in doubly robust estimation. Consistent with the findings in Tables 1 and S1 in the Supplementary Material, the MLE leads to shorter confidence intervals than the DR estimators.

We also compare the regime where all mothers are subject to substantial stress for 8 consecutive days, versus the regime where all mothers are never stressed during these 8 days, possibly due to

a fully effective intervention program. We choose this comparison because it is expected to show the largest effect. We use estimates from the MLE as they are the most stable among the three. Compared to the latter, the former regime is estimated to result in a 4.850 (95% CI [1.202, 8.498]) fold increase in risk of childhood illness on the 9th day, suggesting a statistically significant causal effect by comparing these extreme regimes. Note that, even though we failed to reject the overall g-null in a test with six degrees of freedom, we find statistical significance when comparing this particular pair, for which we anticipate the causal effect to be the largest.

[Table 4 about here.]

7. Discussion

In this paper we introduce a general approach for causal inference from complex longitudinal data with binary outcomes and time-varying confounders. Our approach is based on the SNMMs developed by Robins (1994), which overcome the null paradox of the g-formula and have many important advantages over marginal structural models as detailed in Section 1. Furthermore, SNMMs provide natural non-centrality parameters to describe deviations from the causal g-null. When the outcome is unconstrained, both the multiplicative SNMMs and additive SNMMs are variation independent of the conventional nuisance models as described in Robins (1997, Appendix 2, p.36) and Robins (1994), respectively. However, the conventional multiplicative and additive SNMMs are not suitable for inferring causal effects with binary outcomes as they do not naturally respect the fact that probabilities are bounded above by one, whereas the logistic SNMMs cannot be used in combination with g-estimation methods that are guaranteed to yield a valid test of the g-null in randomized trials. We address this problem in two ways: for binary multiplicative SNMMs we introduce novel nuisance models so that the SNMM parameters can be estimated in a compatible way; for binary additive SNMMs we show that the SNMMs are variation dependent on each other

(and hence incompatible) so that additive SNMMs should probably be avoided when analyzing binary outcomes.

In this article, we have assumed that all the subjects have the same number of follow-ups. In practice, however, it is often the case that study participants receive different number of treatments due to loss to follow-up. In this case, one may first create a pseudo-population by reweighting the observed population with inverse probability of censoring weights, and then apply the proposed estimation methods to the pseudo-population; see Hernán and Robins (2020, §21.5) for a detailed discussion.

Acknowledgements

This research was supported in part by NSERC grants RGPIN-2019-07052, DGECR-2019-00453 and RGPAS-2019-00093, U.S. National Institutes of Health grants R01 AI032475, AI113251, R01AA23187 and P41EB028242 and ONR grants N00014-15-1-2672 and N00014-19-1-2446. The authors thank Robin Evans for helpful comments, and Susan Murphy for kind support. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

Daniel, R. M., Cousens, S., De Stavola, B., Kenward, M. G., and Sterne, J. (2013). Methods for dealing with time-dependent confounding. *Statistics in Medicine*, 32(9):1584–1618.

Henderson, R., Ansell, P., and Alshibani, D. (2010). Regret-regression for optimal dynamic treatment regimes. *Biometrics*, 66(4):1192–1201.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC.

Matsouaka, R. A. and Tchetgen Tchetgen, E. J. (2014). Likelihood based estimation of logistic structural nested mean models with an instrumental variable. Last accessed on Oct 10, 2016.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Qian, T., Yoo, H., Klasnja, P., Almirall, D., and Murphy, S. A. (2021). Estimating time-varying causal excursion effect in mobile health with binary outcomes. *Biometrika*, just-accepted.

Richardson, T. S., Robins, J. M., and Wang, L. (2017). On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association*, pages 1–10.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods*, 23(8):2379–2412.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pages 69–117. Springer.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second Seattle Symposium in Biostatistics*, pages 189–326. Springer.

Robins, J. M., Greenland, S., and Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, 94(447):687–700.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Robins, J. M. and Rotnitzky, A. (2004). Estimation of treatment effects in randomised trials

with non-compliance and a dichotomous outcome using structural mean models. *Biometrika*, 91(4):763–783.

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*, pages 409–420. Morgan Kaufmann Publishers Inc.

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins, 3rd edition.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q-and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science*, 29(4):640.

Shi, C., Fan, A., Song, R., Lu, W., et al. (2018). High-dimensional A-learning for optimal dynamic treatment regimes. *The Annals of Statistics*, 46(3):925–957.

Vansteelandt, S. (2010). Estimation of controlled direct effects on a dichotomous outcome using logistic structural direct effect models. *Biometrika*, 97(4):921–934.

Vansteelandt, S. and Goetghebeur, E. (2003). Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):817–835.

Vansteelandt, S. and Joffe, M. (2014). Structural nested models and g-estimation: The partially realized promise. *Statistical Science*, 29(4):707–731.

Wang, L., Zhang, Y., Richardson, T. S., and Robins, J. M. (2021). Estimation of local treatment effects under the binary instrumental variable model. *Biometrika*.

Yin, J., Markes, S., Richardson, T. S., and Wang, L. (2021). Multiplicative effect modeling: The general case. *Biometrika*, just-accepted.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130.

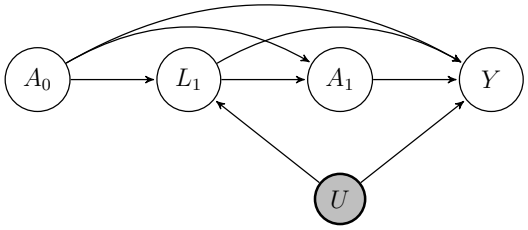

Figure 1: A DAG illustrating time-varying treatments and confounders. The baseline covariates L_0 are omitted for brevity. Variables A_0, L_1, A_1, Y are observed; U is unobserved.

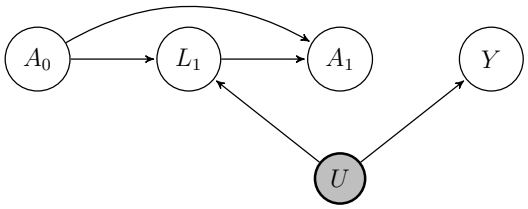

Figure 2: A DAG illustrating the g-null paradox. The baseline covariates L_0 are omitted for brevity. Variables A_0, L_1, A_1, Y are observed; U is unobserved.

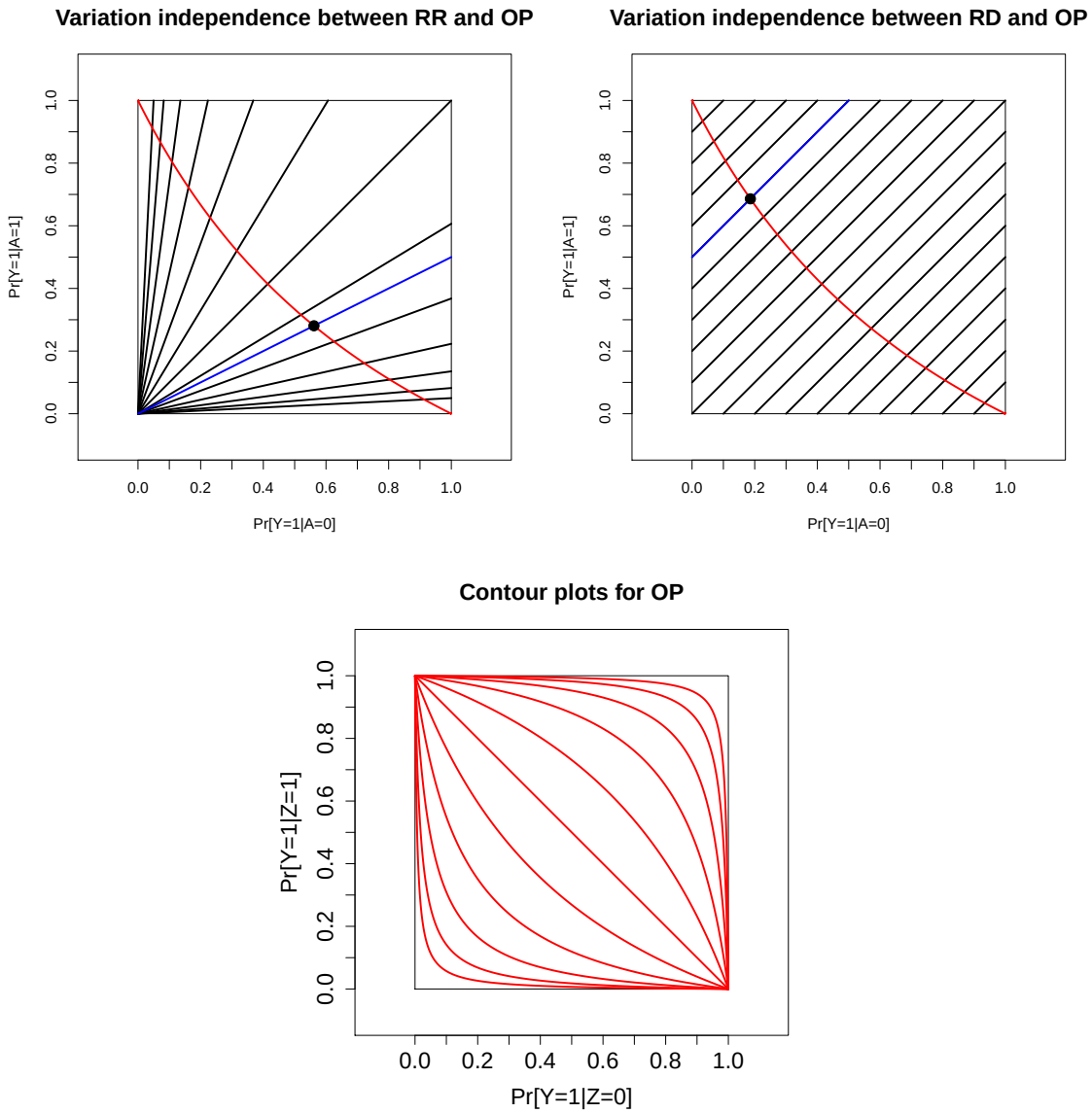

Figure 3: L'Abbé plots: Lines of constant: (Upper left) log RR $\in (-3, -2.5, \ldots, 3)$, OP = 0.5 (red curve); (Upper right) RD $\in \{-0.9, -0.8, \ldots, 0.9\}$, OP = 0.5 (red curve); (Lower) log OP $\in \{-5, -4, \ldots, 5\}$.

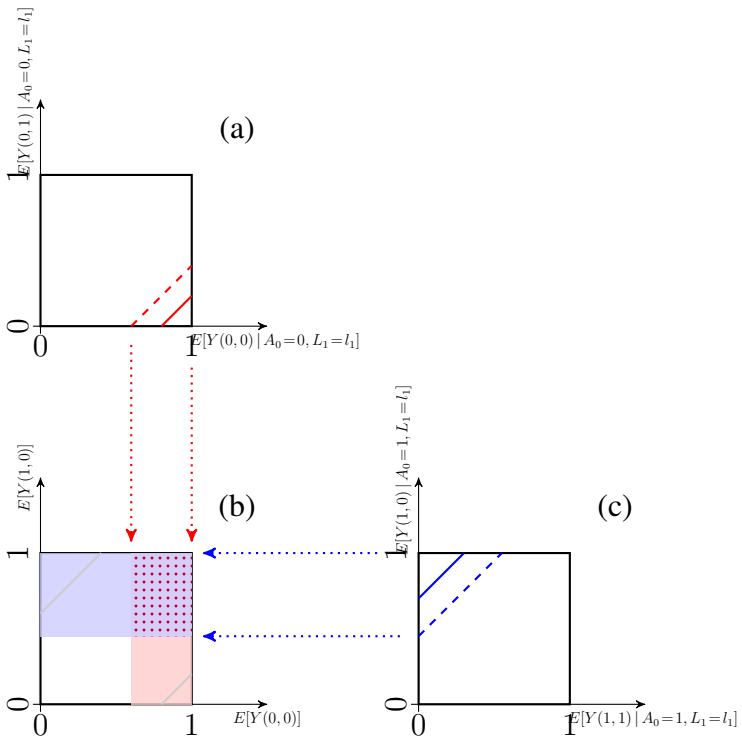

Figure 4: Illustration of variation dependence of additive blip functions. For simplicity, we suppress the dependence on the baseline covariates L_0. The lines at 45 degrees in (a) give values for the second stage additive blip quantities $E[Y(0,1)-Y(0,0) \mid A_0 = 0, L_1 = l_1], l_1 = 0, 1$. Similarly (c) shows the second stage blips with $A_0 = 1$. (b) shows the first stage blip $E[Y(1,0) - Y(0,0)]$. The dotted arrows from (a), (c) to (b) show restrictions placed on the the first stage quantities $E[Y(a_0,0)], a_0 = 0, 1$, due to restrictions on the second stage quantities $E[Y(a_0,0) \mid A_0 = a_0, L_1 = l_1], a_0 = 0, 1$. The dotted area in (b) shows the feasible region of $(E[Y(0,0)], E[Y(1,0)])$ given the second stage quantities depicted in (a) and (c), and the two gray lines at 45 degrees are two contour lines for the first stage blip $E[Y(1,0) - Y(0,0)]$ that are not feasible, i.e. that do no intersect with the dotted region. See the text for more explanation.

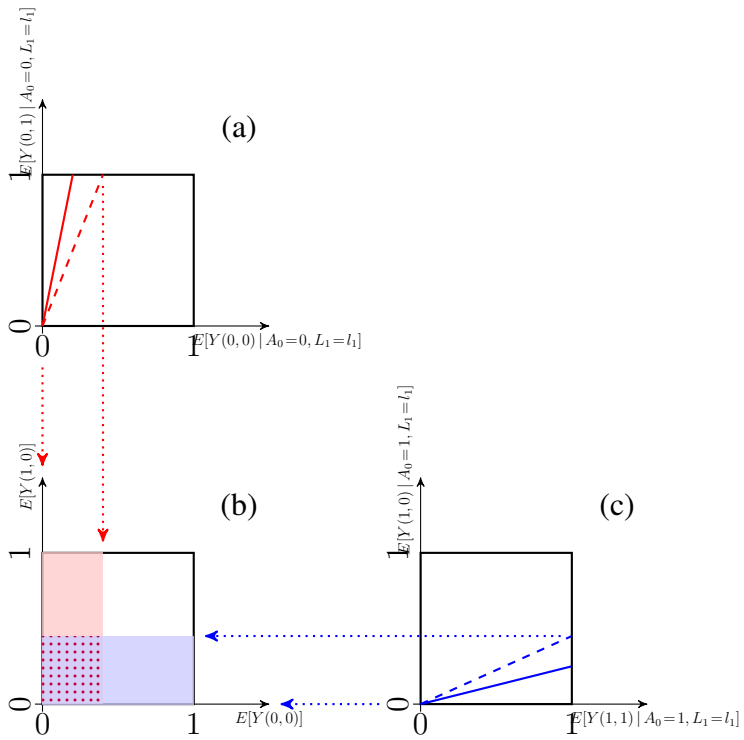

Figure 5: Illustration of variation independence of multiplicative blip functions. For simplicity, we suppress the dependence on the baseline covariates L_0. The lines through the origin in (a) give values for the second stage additive blip quantities $E[Y(0,1) \mid A_0 = 0, L_1 = l_1]/E[Y(0,0) \mid A_0 = 0, L_1 = l_1], l_1 = 0, 1$. Similarly (c) shows the second stage blips with $A_0 = 1$. In (b) the intersection of the shaded regions indicates possible values for $(E[Y(0,0)], E[Y(1,0)])$. Since this set contains the origin $(0,0)$, it will intersect any contour line for the multiplicative first stage blip. Thus the first and second stage multiplicative blips are variation independent.

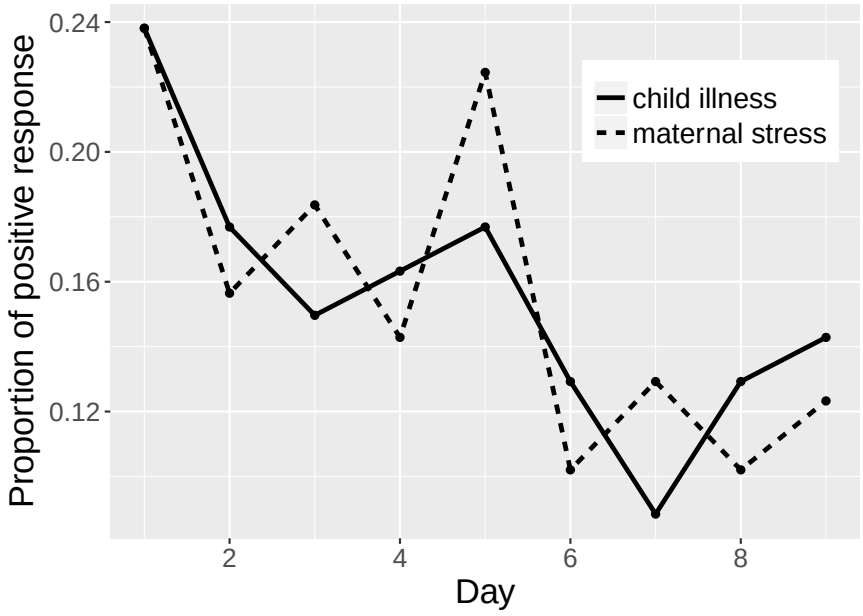

Figure 6: Mothers' stress and children's illness evolving over time.

Table 1: Bias $\times$ 100 (Monte Carlo standard error $\times$ 100) of the proposed methods with a binary baseline covariate L_0. The sample size is 1000

	MLE		2-step MLE		DR	
	baseline	slope	baseline	slope	baseline	slope
SNMM parameters						
$\theta_0(l_0)$	0.63(0.55)	-1.8(0.97)	0.63(0.55)	-1.8(0.98)	0.49(0.55)	1.4(0.97)
$\theta_1(l_0,1,1)$	-1.5(1.2)	3.0(1.6)	-1.5(1.2)	2.9(1.6)	-0.92(1.2)	3.7(1.7)
$\theta_1(l_0,1,0)$	0.80(0.57)	0.60(0.74)	0.79(0.57)	0.68(0.75)	0.87(0.57)	-1.0(0.74)
$\theta_1(l_0,0,1)$	0.90(1.3)	0.15(2.1)	0.91(1.3)	0.045(2.1)	0.75(1.3)	2.7(2.1)
$\theta_1(l_0,0,0)$	0.53(0.60)	-0.40(1.1)	0.53(0.60)	-0.45(1.1)	0.40(0.60)	1.7(1.1)
Nuisance parameters						
$\phi_0(l_0)$	-1.5(1.1)	-1.4(1.9)	-1.5(1.1)	-1.4(1.9)	–	–
$\phi_1(l_0)$	-1.1(0.84)	-0.90(1.3)	-1.1(0.84)	-0.78(1.3)	–	–
gop(l_0)	-6.5(3.5)	2.2(5.5)	-6.5(3.5)	1.4(5.4)	–	–
$\eta_0(l_0)$	-0.73(0.58)	0.73(0.78)	-0.73(0.58)	0.80(0.78)	–	–
$\eta_1(l_0)$	-0.63(0.56)	0.43(0.87)	-0.63(0.56)	0.31(0.87)	–	–
Marginal causal parameters						
$E[Y(0,0)]$	-0.12(0.13)	–	-0.12(0.13)	–	-0.29(0.13)	–
$E[Y(1,1)]$	-0.18(0.14)	–	-0.18(0.14)	–	0.15(0.14)	–
$\frac{E[Y(1,1)]}{E[Y(0,0)]}$	1.3(0.79)	–	1.3(0.79)	–	3.2(0.80)	–

Table 2: Mean computation time in seconds for a Monte Carlo sample with size 1000. The computation time for the DR method does not include the time to get the preliminary estimates and warm starting value using 2-step MLE

	MLE	2-Step MLE	DR
binary L_0	29.31	26.65	4.44
continuous L_0	566.33	514.17	130.99

Table 3: Bias $\times$ 100 (Monte Carlo standard error $\times$ 100) of the proposed DR estimator, and the DR estimator with a logistic baseline model, under misspecification of baseline nuisance models. The propensity score models are correctly specified. The sample size is 1000

	DR (Proposed)		DR (logistic baseline model)	
	baseline	slope	baseline	slope
SNMM parameters				
$\theta_0(l_0)$	0.12(0.54)	2.4(0.97)	2.7(0.51)	-1.5(0.96)
$\theta_1(l_0, 1, 1)$	-0.26(1.2)	2.5(1.7)	-3.0(1.2)	6.9(1.7)
$\theta_1(l_0, 1, 0)$	0.97(0.57)	-1.3(0.74)	0.35(0.56)	-0.48(0.74)
$\theta_1(l_0, 0, 1)$	-0.28(1.3)	5.3(2.1)	24(5.9)	-21(6.7)
$\theta_1(l_0, 0, 0)$	0.26(0.60)	2.1(1.1)	1.2(0.61)	1.1(1.2)

Table 4: Estimation results for the MSCM study

	MLE	DR (Proposed)	DR (logistic baseline model)
p-value for testing the g-null	0.458	0.970	0.779
SNMM coefficients estimates with 95% CI			
α_0	-0.15 (-0.95,0.65)	-2.09 (-7.51,3.32)	2.66 (-4.10,9.42)
α_1	-0.28 (-1.07,0.52)	-1.35 (-6.16,3.47)	2.08 (-5.39,9.55)
α_X			
Household size > 3	0.30 (-0.45,1.04)	0.77 (-3.50,5.03)	0.46 (-3.92,4.85)
Race non-white	0.28 (-0.52,1.08)	-0.03 (-3.88,3.82)	1.15 (-4.24,6.55)
Employed	0.08 (-0.52,0.68)	0.03 (-4.87,4.93)	-3.96 (-10.38,2.46)
Married	-0.15 (-0.67,0.36)	1.68 (-2.20,5.56)	-3.21 (-10.00,3.58)
Computation time*	25.89s	194.67s	193.33s

*: Computation is done using the same resource as in the simulations, and the time reported is the time for getting the point estimate.