Causal Mediation in a Survival Setting with Time-Dependent Mediators

Wenjing Zheng

Division of Biostatistics, University of California, Berkeley

wzheng@stat.berkeley.edu

Mark J. van der Laan

Division of Biostatistics, University of California, Berkeley

laan@berkeley.edu

Abstract

The effect of an expsore on an outcome of interest is often mediated by intermediate variables. The goal of causal mediation analysis is to evaluate the role of these intermediate variables (*mediators*) in the causal effect of the exposure on the outcome. In this paper, we consider causal mediation of a baseline exposure on a survival (or time-to-event) outcome, when the mediator is time-dependent. The challenge in this setting lies in that the event process takes places jointly with the mediator process; in particular, the length of the mediator history depends on the survival time. As a result, we argue that the definition of natural effects in this setting should be based on only blocking those paths from treatment to mediators that are not through the survival history. We propose to use a stochastic interventions (SI) perspective, introduced by Didelez, Dawid, and Geneletti (2006), to formulate the causal mediation analysis problem in this setting. Under this formulation, the mediators are regarded as intervention variables, onto which a given counterfactual distribution is enforced. The natural direct and indirect effects can be defined analogously to the ideas in Pearl (2001). In particular, they also allow for a total effect decomposition and an interpretation of the natural direct effect as a weighted average of controlled direct effects. The statistical parameters that should arise are defined nonparametrically; therefore, they have meaningful interpretations, independent of the causal formulations and assumptions. We present a general semiparametric inference framework for these parameters. Using their efficient influence functions, we develop semiparametric efficient and robust targeted substitution-based (TMLE) and estimating-equation-based (A-IPTW) estimators. An IPTW estimator and g-computation estimator will also be presented.

Keywords: Natural direct effects, natural indirect effects, mediation analysis, mediation formula, mediator, direct effects, causal mediation, survival, time to event, time-dependent, time-varying, robust, double robust, asymptotic linearity, canonical gradient, efficient influence curve, efficient score, loss-based learning, targeted maximum likelihood estimator, targeted learning, parametric working submodels.

1. INTRODUCTION

An exposure often acts on an outcome of interest directly, and/or indirectly through the mediation of some intermediate variables. Identifying and quantifying these two types of effects contribute to further understanding of the underlying causal mechanism. Much of the existing literature on causal mediation is focused on applications in non-survival settings. Causal mediation of a treatment effect on a survival outcome, by contrast, has received relatively fewer attention. In this work, we study mediation analysis in a survival setting with baseline treatment and time-dependent mediator. More specifically, consider a study where covariates and treatment are measured at baseline, and at each follow up visit, one measures the value of an intermediate variable of interest (*mediator*) and whether death or right censoring (e.g. lost to follow up, end of study) has occurred. A subject's observations end after death or right censoring. Suppose we are interested in the effect of the treatment on the time till death (*failure time*), and the mediator lies on the causal pathway between these two — the risk of one dying at a given time depends on the mediator history, which is also affected by the treatment. Therefore, the treatment can act on the failure time directly, and/or indirectly through its effect on the mediator. The goal of mediation analysis is to quantify these two types of treatment effects on the failure time. The challenge in this setting lies in that the outcome of interest is a process (the *event process*) that happens jointly with the mediator process.

One way to assess the direct effect of the treatment on failure time is to compare the distributions of the failure times under different treatments regimens while the mediators are fixed to some common pre-specified values. This is known as the *controlled direct effect* (e.g. Pearl (2001)). Its analysis is very similar to that of a time-dependent deterministic treatment in a non-mediation setting; we refer the reader to existing literature on this topic (e.g. Robins (1997), Hernan, Brumback, and Robins (2000), Stitelman, Gruttola, Wester, and van der Laan (2011)). Controlled direct effects are of interest if the treatment effect under one particular mediator value constitutes a meaningful scientific question. If that is not the case, one may ask a different direct effect question: what would be the effect of treatment on failure time if the treatment had no effect on the mediator (i.e. the mediator takes its value as if treatment were absent)? One way to rigorously formulate this question is using the so-called *natural direct effect* parameter (Robins and Greenland (1992), Pearl (2001)). The natural direct effect has a complementary *natural indirect effect*; together they provide a decomposition of the overall effect of the treatment on the outcome. In this paper, we focus on the natural direct and indirect effects.

In the case of a time-independent mediator that is measured before the onset of the event process (and censoring process), the definition of natural effects and their identifiability (e.g. Lange and Hansen (2011), Tchetgen Tchetgen (2011)) can be extended from the formulations in non-longitudinal setting (e.g. Robins and Greenland (1992), Pearl (2001), Robins (2003), Petersen, Sinisi, and van der Laan (2006), Imai, Keele, and Yamamoto (2010)). For inference of these parameters, the use of additive hazard models for the outcome and linear models for the mediator are proposed in Lange and Hansen (2011); the use of accelerated failure time and proportional hazard models are studied in Tein and MacKinnon (2003) and VanderWeele (2011); robust estimators for the natural direct and indirect effects under proportional hazards models and additive hazards models, as well as sensitivity analysis techniques for assessing the impact of the violation of the mediator's ignorability assumption, are developed in Tchetgen Tchetgen (2011). A more complex variation of this setting is when there is a confounder (the *recanting witness* covariate, Avin, Shpitser, and Pearl

(2005)) of the mediator–outcome relation that is affected by the treatment of interest. Robins and Richardson (2010) show that with additional conditions regarding independence (or deterministic dependence) of the counterfactual recanting witness under various treatment levels, the resulting mediation parameters can be identified. Tchetgen Tchetgen and VanderWeele (2012) show that these additional conditions can be avoided in some special cases of binary recanting witness, or under additional parametric assumptions on the mediator–recanting witness relation.

In the case of a time-dependent mediator, the probability of the mediator process having a given non-degenerate length would depend on the failure time. This interdependency between the event process and the mediator process poses a challenge when extending the results from the time-independent mediator setting. Firstly, the event history affects both the current mediator (taking a non-degenerate value) and the current event indicator, but it is part of the outcome of interest and thus is not a recanting witness covariate. More specifically, the treatment affects a current mediator both directly and indirectly through its effect on the event history. In asking what is the effect of a given treatment level on the event process not mediated by the mediator process, one must specify how the paths from treatment to mediator should be blocked. If one blocks all paths from treatment to mediators (both the direct paths of treatment to mediator and the paths through event history), then the parameters defined would be a direct generalization of the definition of mediation formula and natural effects in time-independent mediator settings (by regarding event history as a recanting witness). However, this generalization would yield parameters that are not interpretable for the purpose of effect mediation in this survival setting, since the relation of treatment and event process (outcome of interest) is also altered; we elaborate this further in appendix A4.1. In this light, we argue that, for the current survival setting, the definition of mediation formula and natural effects should be based on blocking only those paths from the treatment to mediator that are not through survival history (these would be an extension of the path-specific effects discussed in Pearl (2001), Avin et al. (2005), Robins and Richardson (2010)). The direct effect question these parameters would address is: what is the effect of treatment on the survival time, if the treatment had no effect on the mediator process other than through survival history?

Having specified the effects of interest, the second challenge arises in formulating these as parameters in a causal framework. Under the traditional definition of mediation parameters in a non-longitudinal setting (e.g. Robins and Greenland (1992), Pearl (2001), Avin et al. (2005), Robins and Richardson (2010)), the mediator is regarded as intermediate counterfactual outcome. In extending this definition to our effects of interest in the current setting, the time-varying mediators become intermediate counterfactual outcomes under a different treatment level than that of their parent counterfactual survival history. Consequently, the identifiability conditions of the resulting parameters would impose restrictions on the event indicators — conditions that we find too strong for the purpose of a survival study (this causal formulation is elaborated in detail in appendix A4.2, the resulting statistical parameters are the same as those in the main text). As an alternative, we propose to adopt a stochastic interventions (SI) perspective to causal mediation, introduced by Didelez et al. (2006). Under this formulation, the mediators are regarded as intervention variables, onto which a given counterfactual distribution is enforced. The natural effects can be defined analogously to the ideas in Pearl (2001) and Avin et al. (2005). In particular, they also allow for a total effect decomposition and an interpretation of the natural direct effect as a weighted average of controlled direct effects. Importantly, however, one should note that even though these SI-based parameters and their non-SI-based counterparts in appendix A4.2 all identify to the same statistical parameters, they are formally different causal parameters defined under different formulations (but aim to answer the same type of mediation questions). For concreteness, we will use the probability of surviving beyond a given time as the effect measure of interest.

The statistical parameters that should arise have meaningful interpretations, regardless of the causal formulations and assumptions; we develop a general semiparametric inference framework for these parameters. More specifically, we will derive the efficient influence functions under a locally saturated semiparametric model, and establish their robustness properties. The variances of these functions provide local efficiency bounds, and their robustness properties give information on the types of model mis-specifications that would still allow for unbiased estimation of the parameters. These efficient influence functions can be used to construct robust and locally semiparametric efficient estimators (e.g. an estimating-equation-based A-IPTW estimator, or a substitution-based TMLE estimator).

This paper proceeds as follows. We begin by considering the case with no right censoring (section 2), as it allows us to focus on the mediator–outcome relation; we then generalize the results to the case with right censoring (section 3). In section 2.1, we define the causal parameters of interest and establish their identifiability conditions. Separately, non-SI based formulations are discussed in appendix A4. Section 2.2 concerns the semiparametric inference of the statistical parameters. The efficient influence curves are derived in section 2.2.1, and their robustness properties are studied in section 2.2.2. In section 2.3, we present the g-computation, IPTW, A-IPTW and TMLE estimators for the natural direct effect; a simulation study (section 2.3.5) is conducted to evaluate their performances. Similar estimators for the natural indirect effect are given in appendix A3. In section 3, we extend the results to the case with right censoring. We will only focus on the identification and the efficient influence function of the mediation formula. The corresponding results for the natural direct and indirect effects can be derived from these by following the steps in section 2. The paper concludes with a discussion section.

2. NO RIGHT CENSORING.

For simplicity, we begin by considering the case when there is no right censoring (i.e. failure times are always observed). Firstly, we establish the definition of a counterfactual failure time pertinent to mediation analysis in the present setting, and determine the identifiability conditions for the parameters of interest. Thereafter, we derive the efficient influence functions of the parameters under a locally saturated semiparametric model, and present the g-computation, IPTW, A-IPTW and TMLE estimators for these parameters. Generalization of these results to account for right censoring are addressed in section 3.

2.1 Data and parameters of interest

Consider a study where each individual's baseline covariates $W \in \mathscr{W}$ and a baseline treatment $A \in \mathscr{A}$ are measured at the beginning of the study $(t = 0)$. At each of the subsequent follow-up visit $t \in \{1, \ldots, \tau\}$, one measures the value of the mediator $Z_t \in \mathscr{Z}$, and whether death (or the event of interest) has occurred. Let T denote the visit where death was first reported. We refer to T as the *failure time*. In this section, we assume that T is always observed. The observation on an individual is given by $O = (W, A, (Z_1, \ldots, Z_T), T)$, since the records end after death (or event of interest). Let $N_t \equiv I(T \leq t)$ be a process that jumps to 1 after death, and let $dN_t \equiv I(T = t)$ denote the event indicator. The data structure can be represented as $O =$

$(W,A,(Z_t,dN_t : t = 1,\ldots,\tau))$, where for $t > T$, Z_t is given a degenerate value that is outside of $\mathscr{Z}$. Let P_0 denote the probability distribution of O. The observed data consists of n i.i.d observations of $O \sim P_0$.

From here on, for any $1 \leq t \leq \tau$ and a time-dependent variable V, we will use the boldface $\mathbf{V}_t$ to denote the vector $(V_1,\ldots,V_t)$, use $\mathbf{V}_{j \geq t}$ to denote the vector $(V_t,\ldots,V_\tau)$. When referring to the entire vector $\mathbf{V}_\tau$, we will also use the shorthand $\mathbf{V}$. For any $1 \leq s \leq t \leq \tau$, $\mathbf{V}_s^t$ will denote $(V_s,\ldots,V_t)$. Degenerate indices such as $\mathbf{V}_{-1}$ or $\mathbf{V}_s^{s-1}$ all signify the empty set.

The following Non-Parametric Structural Equations Model (NPSEM, Pearl (2009)) encodes the time-ordering assumption on the variables:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W,U_A) \\ Z_t &= f_{Z_t}(W,A,\mathbf{Z}_{t-1},N_{t-1},U_{Z_t}) \text{ for } t = 1,\ldots,\tau \\ dN_t &= f_{dN_t}(W,A,\mathbf{Z}_t,N_{t-1},U_{dN_t}) \text{ for } t = 1,\ldots,\tau. \end{aligned} \tag{1}$$

where $U \equiv (U_W,U_A,(U_{Z_t},U_{dN_t} : t))$ is the set of unobserved exogenous variables, and $X \equiv (W,A,\mathbf{Z},\mathbf{dN})$ is the set of endogenous variables, and $\{f_{X_j} : j\}$ are unspecified deterministic functions. Counterfactuals under the Rubin Causal Model (e.g. Rubin (1978), Rosenbaum and Rubin (1983) and Holland (1986)) can be represented as restrictions on the input of the functions f_{X_j}. For instance, given a and $\mathbf{z}$, $dN_t(a,\mathbf{z}) \equiv f_{dN_t}(W,A=a,\mathbf{Z}_t=\mathbf{z}_t,N_{t-1}(a,\mathbf{z}),U_{dN_t})$ corresponds to the event process that the individual would have followed if, all else equal, he/she had treatment $A=a$ and mediator process $\mathbf{Z}=\mathbf{z}$. More rigorously, such a counterfactual process ensue from a hypothetical experiment that first measures the pre-treatment covariates W, then sets the treatment to $A = a$, and, at each subsequent time, sets the mediator to take value

$Z_t = z_t$ and records the resulting event indicator.

The observed data structure is generated from (1) without any interventions; in other words, $O = (W, A, \mathbf{Z}(A), \mathbf{dN}(A))$. The likelihood of $O \sim P_0$ can be factored according to the time-ordering:

$$p_0(O) = p_0(W)p_0(A \mid W)\prod_{t=1}^{\tau} p_0(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1})p_0(dN_t \mid W, A, \mathbf{Z}_t, N_{t-1}). \qquad (2)$$

Recall that if the event occurred at T, then for all $t > T$, Z_t are assigned a degenerate value with probability 1, and $dN_t = 0$ with probability 1. We adopt the notations $g_{A,0}(A \mid W) \equiv p_0(A \mid W)$, $g_{Z,0}(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1}) \equiv p_0(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1})$, $g_0 \equiv (g_{A,0}, g_{Z,0})$, $Q_{W,0}(W) \equiv p_0(W)$, $Q_{dN,0}(t \mid W, A, \mathbf{Z}_t, N_{t-1}) \equiv p_0(dN_t = 1 \mid W, A, \mathbf{Z}_t, N_{t-1})$, and $Q_0 \equiv (Q_{W,0}, Q_{dN,0})$.

2.1.1. Counterfactual failure time

As argued in the introduction, since the treatment affects the mediators both directly and through its effect on the survival history of interest, the mediation parameters should be defined based on blocking only those paths from treatment to mediator that are not through the survival history (see appendix A4 for more details).

To define the pertinent counterfactual failure time, we propose to use a stochastic interventions (SI) perspective introduced by Didelez et al. (2006). Stochastic interventions (a.k.a stochastic policies, random interventions, randomized dynamic strategies; e.g. Dawid and Didelez (2010), Pearl (2009), Tian (2008), Robins and Richardson (2010), Diaz and van der Laan (2011)) are generalizations of the traditional static interventions or dynamic regimes where, instead of assigning a deterministic value, one assigns a probability distribution to an intervention variable. Using the non-longitudinal setting as background, Didelez et al. (2006) illustrate how the notions of various direct and indirect effects can be formulated as sequential treatment problems by regarding the mediator as an intervention variable (receiving either a deterministic or stochastic intervention). Though their approach was based on a non-counterfactual causal framework (e.g. Dawid and Didelez (2010)), the essence of their idea remains the same, and is easily generalizable to survival settings.

Let a and a' be two possible treatment levels. Let $Z_t(a')$ and $dN_t(a')$ denote the counterfactual mediator and event indicator under an intervention which sets $A = a'$. Let $g_{Z(a')}$ denote the conditional distribution of $Z_t(a')$, i.e., $g_{Z(a')}(z_t \mid w, \mathbf{z}_{t-1}, n_{t-1}) \equiv p(Z_t(a') = z_t \mid W = w, \mathbf{Z}_{t-1}(a') = \mathbf{z}_{t-1}, N_{t-1}(a') = n_{t-1})$. Consider an intervention which imposes the following conditional distribution $g_{a,a'}$ on $(A, Z_1, \ldots, Z_\tau)$:

$$\begin{aligned} &g_{a,a'}(A = a \mid W) \equiv 1 \\ &g_{a,a'}(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1}) \equiv g_{Z(a')}(Z_t \mid W, \mathbf{Z}_{t-1}, N_{t-1}). \end{aligned} \tag{3}$$

The resulting counterfactual event process is $dN_t(a, Z(g_{a,a'}))$, denote the resulting failure time by $T(a, Z(g_{a,a'}))$. This experiment is encoded as:

$$\begin{aligned} W &= f_W(U_W) \\ A &= a \\ Z_t(g_{a,a'}) &= f^{a'}_{Z_t}(W, A = a', \mathbf{Z}_{t-1}(g_{a,a'}), N_{t-1}(a, Z(g_{a,a'})), U^{a'}_{Z_t}), \text{ for } t = 1, \ldots, \tau \\ dN_t(a, Z(g_{a,a'})) &= f_{dN_t}(W, A = a, \mathbf{Z}_t(g_{a,a'}), N_{t-1}(a, Z(g_{a,a'})), U_{dN_t}), \text{ for } t = 1, \ldots, \tau. \end{aligned} \tag{4}$$

In words, this experiment first sets the baseline treatment to $A = a$. Then, at each visit t, for a given realization of $(W, A = a, \mathbf{Z}_{t-1}(g_{a,a'}), N_{t-1}(a, Z(g_{a,a'}))) = (w, a, \mathbf{z}_{t-1}, n_{t-1})$, it sets $Z_t(g_{a,a'})$ to be distributed according to

$$P(Z_t(g_{a,a'}) = \cdot \mid W = w, A = a, \mathbf{Z}_{t-1}(g_{a,a'}) = \mathbf{z}_{t-1}, N_{t-1}(a, Z(g_{a,a'})) = n_{t-1}) \equiv g_{Z(a')}(\cdot \mid w, \mathbf{z}_{t-1}, n_{t-1})$$

(recall that if death has already occurred, i.e. $n_{t-1} = 1$, then $g_{Z(a')}$ will assign the degenerate value with probability 1); it then measures the response $dN_t(a, Z(g_{a,a'}))$ under realized history $(W = w, A = a, \mathbf{Z}_t(g_{a,a'}) = \mathbf{z}_t, N_{t-1}(a, Z(g_{a,a'})) = n_{t-1})$. The joint distribution of

$$\left(U \equiv (U_W, U_A, \mathbf{U}_Z \equiv (U_{Z_t} : t), \mathbf{U}_{dN} \equiv (U_{dN_t} : t)), U^g_Z \equiv \{(U^{a'}_{Z_t} : t) : a' \in \mathscr{A}\}\right),$$

together with the structural equations $f_X \equiv (f_W, f_A, (f_{Z_t} : t), (f_{dN_t} : t), \{(f^{a'}_{Z_t} : t) : a' \in \mathscr{A}\})$, define a full data random variable $(U, U^g_Z) \sim P_{(U, U^g_Z)}$ on an individual. Counterfactual variables that arise are subset of this full random variable.

Note that each $Z_t(g_{a,a'})$ is an intervention variable which, given a realized history, follows the specified intervention distribution $g_{Z(a')}(\cdot \mid w, \mathbf{z}_{t-1}, n_{t-1})$; this is different from a counterfactual response variable $Z_t(a')$ under intervention $A = a'$. For the former, the event history affecting $Z_t(g_{a,a'})$ is $N_{t-1}(a, Z(g_{a,a'}))$ under treatment $A = a$; for the latter, the event history affecting $Z_t(a')$ is $N_{t-1}(a')$ under treatment $A = a'$. In appendix A4.1, we study the experiment defined using the latter.

As mentioned earlier, even if one can carry out an intervention on the mediator (separately from the intervened treatment), the SI formulation formally requires the external specification of the function $g_{Z(a')}(z_t \mid w, \mathbf{z}_{t-1}, n_{t-1})$, which is the conditional distribution of the counterfactual variable $Z_t(a')$. If this conditional distribution is not known, it needs to be ascertained using a separate controlled experiment which sets $A = a'$ throughout (1). Therefore, aside from causal assumptions needed to identify the distribution of $dN_t(a, Z(g_{a,a'}))$ (for given $g_{Z(a')}$) in the main experiment, additional assumptions are needed to identify $g_{Z(a')}$ as a function of the data generating distribution (see next section).

2.1.2. Mediation formula, natural direct and indirect effects

For concreteness, suppose one is interested in the effect of a binary treatment on the probability of the patient surviving beyond a specific time t_0. We refer to the difference $P(T(1, Z(g_{1,1})) > t_0) - P(T(1, Z(g_{1,0})) > t_0)$ as the *natural indirect effect* (NIE) and the difference $P(T(1, Z(g_{1,0})) > t_0) - P(T(0, Z(g_{0,0})) > t_0)$ as the *natural direct effect* (NDE). The identification and estimation of these two effects can be approached through the so-called *mediation formula* (Pearl (2011)):

$$\Psi_{a,a'}(P_{(U,U_Z^g)}) \equiv P(T(a,Z(g_{a,a'})) > t_0). \tag{5}$$

It is important to note that while the definition of these parameters are analogous to those in Robins and Greenland (1992), Pearl (2001), Pearl (2011) and Avin et al. (2005), they are ultimately not the same definitions since that the mediator variables are conceptualized differently. In this respect, the parameters defined here aim to provide alternative formulations to questions that arise in mediation analysis in the current survival setting.

The identifiability of these parameters is a consequence of established results regarding stochastic interventions (e.g. Dawid and Didelez (2010), Pearl (2009), Robins and Richardson (2010)).

Theorem 1. *Suppose the following positivity assumptions regarding the data generating distribution P_0 hold:*

P1. *There exists $0 < \delta_A < 1$ such that $g_{A,0}(A \mid W) > \delta_A$, a.e. over $\mathscr{A}$;*

P2. *There exists $0 < \delta_Z < 1$ such that $\inf_{t \in \{1,\ldots,t_0\}} g_{Z,0}(Z_t \mid W, A = a, \mathbf{Z}_{t-1}, N_{t-1} = 0) > \delta_Z$, a.e. over $\mathscr{Z}$;*

P3. *There exists $0 < \delta_N < 1$ such that $\inf_{t \in \{1,\ldots,t_0\}} 1 - Q_{dN,0}(t \mid W, A = a', \mathbf{Z}_t, N_{t-1} = 0) > \delta_N$.*

Let $g_{a,a'}$ be an intervention distribution on $(A,\mathbf{Z})$ as defined in (3). Let $T(a,Z(g_{a,a'}))$ be the corresponding failure time under experiment (4). Suppose the following randomization assumptions hold for all $\mathbf{z}$:

I1. *$(\mathbf{dN}(a'), \mathbf{Z}(a')) \perp A$, given W.*

I2. *$\mathbf{dN}(a,\mathbf{z}) \perp A$, given W;*

I3. $\mathbf{dN}_{j\geq t}(a,\mathbf{z}) \perp Z_t$*, given* $W, A = a, \mathbf{Z}_{t-1} = \mathbf{z}_{t-1}, N_{t-1}$*;*

I4. $\mathbf{dN}_{j\geq t}(a,\mathbf{z}) \perp Z_t(g_{a,a'})$*, given* W*,* $A = a$*,* $\mathbf{Z}_{t-1}(g_{a,a'}) = \mathbf{z}_{t-1}$*,* $N_{t-1}(a, Z(g_{a,a'}))$*;*

then, (5) can be expressed as

$$\Psi_{a,a'}(P_0) \equiv$$
$$\sum_{w\in\mathscr{W}} \sum_{\mathbf{z}_{t_0}\in\mathscr{Z}^{t_0}} Q_{W,0}(w) \prod_{t=1}^{t_0} \left\{ g_{Z,0}\left(z_t \mid w, A = a', \mathbf{z}_{t-1}, N_{t-1} = 0\right) \left(1 - Q_{dN,0}(t \mid w, A = a, \mathbf{z}_t, N_{t-1} = 0)\right) \right\}, \quad (6)$$

where $\mathscr{Z}^{t_0} \equiv \prod_1^{t_0} \mathscr{Z}$ *is the outcome space of the vector* $\mathbf{Z}_{t_0}$*.*

Proof. See Appendix A2 □

Note that when $a = a'$, (6) indeed equals the g-computation formula for the parameter $P(T(a) > t_0)$. This ensures that the g-computation formula of $P(T(1) > t_0) - P(T(0) > t_0)$ decomposes into the g-computation formulas for NIE (9) and NDE (8) below. The parameter (6) also equals the g-computation formula for path-specific effects discussed in Avin et al. (2005) and Robins and Richardson (2010). In fact, the non-SI based parameters we consider in appendix A4.2 would have g-computation formulas (6), (8) and (9).

In words: Conditions I1 and I2 require randomization of the baseline treatment. Condition I3 requires the mediators $\mathbf{Z}$ are sequentially randomized in the observed data. It is important to note that, unlike treatment assignment, mediator variables may not always be amenable to randomization in practice. Condition I4 requires that $\mathbf{Z}$ are sequentially randomized in the hypothetical experiment $g_{a,a'}$; in other words, each variable Z_t under distribution given by $g_{Z(a')}$ needs to be conditionally independent of future potential outcomes.

The parameter (5) can also be identified under conditional independence conditions on the joint distribution of (U, U^g): $(\mathbf{U}_Z, \mathbf{U}_{dN}) \perp U_A$ given U_W; $(\mathbf{U}_{dN})_{j\geq t} \perp U_{Z_t}$ given $U_W, U_A, (\mathbf{U}_Z)_{t-1}, (\mathbf{U}_{dN})_{t-1}$; $(\mathbf{U}_{dN})_{j\geq t} \perp U^{a'}_{Z_t}$, given U_W, U_A, $(\mathbf{U}_Z)^{a'}_{t-1}$, $(\mathbf{U}_{dN})_{t-1}$.

For the rest of this paper, we will suppress the outcome spaces in the subscript for the sums. It should be understood that $\sum_w$ means $\sum_{w\in\mathscr{W}}$ and $\sum_{\mathbf{z}_{t_0}}$ means $\sum_{\mathbf{z}_{t_0}\in\mathscr{Z}^{t_0}}$, unless otherwise noted. From here onward, we will adopt the notations

$$
\begin{aligned}
G_{Z,0}(\mathbf{Z}_t \mid W,A)) &\equiv \prod_{t'=1}^{t} g_{Z,0}\left(Z_{t'} \mid W,A,\mathbf{Z}_{t'-1},N_{t'-1}=0\right) \\
\bar{Q}_{N,0}(t \mid W,A,\mathbf{Z}_t) &\equiv \prod_{t'=1}^{t} 1-Q_{dN,0}(t' \mid W,A,\mathbf{Z}_{t'},N_{t'-1}=0). \qquad (7)
\end{aligned}
$$

Natural direct and indirect effects are functions of the mediation formula, and hence can be identified under the same conditions. More specifically, if the assumptions in theorem 1 hold for $a,a'\in\{0,1\}$, then the natural direct effect $\Psi_{NDE}(P_{(U,U_Z^g)})\equiv P(T(1,Z(g_{1,0}))>t_0)-P(T(0,Z(g_{0,0}))>t_0)$ can be expressed as

$$
\begin{aligned}
&\Psi_{NDE}(P_0) \\
&\equiv \sum_w \sum_{\mathbf{z}_{t_0}} Q_{W,0}(w)\left\{\prod_{t=1}^{t_0} g_{Z,0}\big(z_t \mid w,A=0,\mathbf{z}_{t-1},N_{t-1}=0\big)\big(1-Q_{dN,0}(t \mid w,A=1,\mathbf{z}_t,N_{t-1}=0)\big)\right. \\
&\left.-\prod_{t=1}^{t_0} g_{Z,0}\big(z_t \mid w,A=0,\mathbf{z}_{t-1},N_{t-1}=0\big)\big(1-Q_{dN,0}(t \mid w,A=0,\mathbf{z}_t,N_{t-1}=0)\big)\right\} \\
&= E_{Q_{W,0}}\left\{\sum_{\mathbf{z}_{t_0}} G_{Z,0}(\mathbf{z}_{t_0} \mid W,A=0)\big(\bar{Q}_{N,0}(t_0 \mid W,A=1,\mathbf{z}_{t_0})-\bar{Q}_{N,0}(t_0 \mid W,A=0,\mathbf{z}_{t_0})\big)\right\}, \qquad (8)
\end{aligned}
$$

and the natural indirect effect, defined as $\Psi_{NIE}(P_{(U,U_Z^g)})\equiv P(T(1,Z(g_{1,1}))>t_0)-P(T(1,Z(g_{1,0}))>t_0)$, can be expressed as

$$
\begin{aligned}
&\Psi_{NIE}(P_0) \\
&\equiv \sum_w \sum_{\mathbf{z}_{t_0}} Q_{W,0}(w)\left\{\prod_{t=1}^{t_0} g_{Z,0}\big(z_t \mid w,A=1,\mathbf{z}_{t-1},N_{t-1}=0\big)\big(1-Q_{dN,0}(t \mid w,A=1,\mathbf{z}_t,N_{t-1}=0)\big)\right. \\
&\left.-\prod_{t=1}^{t_0} g_{Z,0}\big(z_t \mid w,A=0,\mathbf{z}_{t-1},N_{t-1}=0\big)\big(1-Q_{dN,0}(t \mid w,A=1,\mathbf{z}_t,N_{t-1}=0)\big)\right\} \\
&= E_{Q_{W,0}}\left\{\sum_{\mathbf{z}_{t_0}}\big(G_{Z,0}(\mathbf{z}_{t_0} \mid W,A=1)-G_{Z,0}(\mathbf{z}_{t_0} \mid W,A=0)\big)\bar{Q}_{N,0}(t_0 \mid W,A=1,\mathbf{z}_{t_0})\right\}, \qquad (9)
\end{aligned}
$$

Note that under positivity and conditions I1-I3 alone, the statistical parameter (8) equals the population mean of a weighted average of controlled direct effects (CDE): $E_W\left(\prod_{t=1}^{t_0} g_{Z(0)}(z_t \mid W, \mathbf{z}_{t-1}, n_{t-1}=0)\right)\left(P(T(1,\mathbf{z})>t_0 \mid W) - P(T(0,\mathbf{z})>t_0 \mid W)\right)$. With condition I4, this weighted CDE equals the natural direct effect $P(T(1,Z(g_{1,0}))>t_0) - P(T(0,Z(g_{0,0}))>t_0)$. Moreover, the total effect can be decomposed as the sum of two effects:

$$\begin{aligned}
&P(T(1)>t_0) - P(T(0)>t_0)\\
&= \left(P(T(1,Z(g_{1,1}))>t_0) - P(T(1,Z(g_{1,0}))>t_0)\right)\\
&+ \left(P(T(1,Z(g_{1,0}))>t_0) - P(T(0,Z(g_{0,0}))>t_0)\right).
\end{aligned}$$

Now that we have identified the statistical parameters of interest, the remaining of this section will concern their inference.

2.2 Semiparametric inference

The statistical parameters of interest (6), (8) and (9) have meaningful interpretations under positivity and time-ordering assumptions alone, regardless of the causal formulations and assumptions. In this section, we develop a general semiparametric inference framework for these parameters. In particular, we derive the Efficient Influence Functions (EIF) of (6), (8) and (9) under a (locally saturated) semiparametric model, and establish their robustness properties. For a given pathwise-differentiable parameter Ψ, under certain regularity conditions, the variance of the EIF of Ψ is a generalized Cramer-Rao lower bound for the variances of the influence functions of asymptotically linear estimators of Ψ. Therefore, the variance of the EIF provides an efficiency bound for the regular and asymptotically linear (RAL) estimators of Ψ. Moreover, under a locally saturated model, the influence function of any RAL estimator is in fact the EIF. We refer the reader to Bickel, Klaassen, Ritov, and Wellner

(1997) for general theory of efficient semiparametric inference.

2.2.1. Efficient influence functions

Let $\mathscr{M}$ denote a locally saturated semiparametric model containing the true data generating distribution P_0. Let P_n denote the empirical distribution of n i.i.d observations of $O \sim P_0$. For a function $f(O)$, we will use Pf to denote the expectation of $f(O)$ under the probability distribution $P \in \mathscr{M}$.

For any $P \in \mathscr{M}$, the likelihood can be factorized according to the time-ordering. We also adopt for P the notations: $g_A(A \mid W) \equiv p(A \mid W)$, $g_Z(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1}) \equiv p(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1})$, $g \equiv (g_A, g_Z)$, $Q_W(W) \equiv p(W)$, $Q_{dN}(t \mid W, A, \mathbf{Z}_t, N_{t-1}) \equiv p(dN_t = 1 \mid W, A, \mathbf{Z}_t, N_{t-1})$, and $Q \equiv (Q_W, Q_{dN})$. This way, P is represented by its components (g, Q). The shorthand notations $G_Z(\mathbf{Z}_t \mid W, A)$ and $\bar{Q}_N(t \mid W, A, \mathbf{Z}_t)$ are defined similarly to (7).

The mediation formula in (6) can be considered as the following map evaluated at P_0:

$$\begin{aligned}\Psi_{a,a'} :& \mathscr{M} \to \mathbb{R} \\ & P \mapsto \Psi_{a,a'}(P) \equiv \sum_{w} \sum_{\mathbf{z}_{t_0}} Q_W(w) G_Z(\mathbf{z}_{t_0} \mid w, A = a') \bar{Q}_N(t_0 \mid w, A = a, \mathbf{z}_{t_0}). \qquad (10)\end{aligned}$$

For later convenience, let us refer to each inner expectation with respect to Z_t in (10) as the *conditional mediation formula at t*:

$$\begin{aligned}&\phi_{Z,a,a'}(P)(t; W, \mathbf{Z}_{t-1}) \\ &\equiv \sum_{\mathbf{z}_t^{t_0}} \prod_{t'=t}^{t_0} g_Z(z_{t'} \mid W, A = a', \mathbf{Z}_{t-1}, \mathbf{z}_t^{t'-1}, N_{t'-1} = 0) \bar{Q}_N\left(t_0 \mid W, A = a, \mathbf{Z}_{t-1}, \mathbf{z}_t^{t_0}\right),\end{aligned}$$

for $t = 1, \ldots, t_0$, and

$$\phi_{Z,a,a'}(P)(t_0 + 1; W, \mathbf{Z}_{t_0}) \equiv \bar{Q}_N(t_0 \mid W, A = a, \mathbf{Z}_{t_0}).$$

Recall that $\mathbf{z}_t^{t_0} \equiv (z_t, \ldots, z_{t_0})$; the sum above is taken over the outcome space $\prod_t^{t_0} \mathscr{Z}$ of $\mathbf{Z}_t^{t_0}$. Note that $\Psi_{a,a'}(P) = E_{Q_W}\left(\phi_{Z,a,a'}(P)(t=1;W)\right)$, and $\phi_{a,a'}(P)$ satisfies the recursive relation

$$\phi_{Z,a,a'}(P)(t-1;W,\mathbf{Z}_{t-2}) = E_{g_{Z,t-1}}\left(\phi_{Z,a,a'}(P)(t;W,\mathbf{Z}_{t-1}) \mid W,A=a',\mathbf{Z}_{t-2},N_{t-2}=0\right). \quad (11)$$

That is, $\phi_{Z,a,a'}(P)(t-1;W,\mathbf{Z}_{t-2})$ is the conditional expectation of $\phi_{Z,a,a'}(P)(t;W,\mathbf{Z}_{t-1})$, conditioned on $(W,A=a',\mathbf{Z}_{t-2},N_{t-2}=0)$, under the mediator distribution g_Z of P.

Similarly, the natural direct effect in (8) and the natural indirect effect in (9) are, respectively, the following maps evaluated at P_0:

$$\begin{aligned} &P \mapsto \Psi_{NDE}(P) \\ &\equiv \sum_w \sum_{\mathbf{z}_{t_0}} Q_W(w) G_Z(\mathbf{z}_{t_0} \mid w, A=0) \left(\bar{Q}_N(t_0 \mid w, A=1, \mathbf{z}_{t_0}) - \bar{Q}_N(t_0 \mid w, A=0, \mathbf{z}_{t_0})\right), \end{aligned} \quad (12)$$

and

$$\begin{aligned} &P \mapsto \Psi_{NIE}(P) \\ &\equiv \sum_w \sum_{\mathbf{z}_{t_0}} Q_W(w) \left(G_Z(\mathbf{z}_{t_0} \mid w, A=1) - G_Z(\mathbf{z}_{t_0} \mid w, A=0)\right) \bar{Q}_N(t_0 \mid w, A=1, \mathbf{z}_{t_0}). \end{aligned} \quad (13)$$

We also adopt the definition of the *conditional natural direct effect at t*:

$$\begin{aligned} &\phi_{Z,NDE}(P)(t;W,\mathbf{Z}_{t-1}) \equiv \phi_{Z,1,0}(P)(t;W,\mathbf{Z}_{t-1}) - \phi_{Z,0,0}(P)(t;W,\mathbf{Z}_{t-1}) \\ &= \sum_{\mathbf{z}_t^{t_0}} \left\{ \prod_{t'=t}^{t_0} g_Z(z_{t'} \mid W, A=0, \mathbf{Z}_{t-1}, \mathbf{z}_t^{t'-1}, N_{t'-1}=0) \right. \\ &\left. \times \left(\bar{Q}_N\left(t_0 \mid W, A=1, \mathbf{Z}_{t-1}, \mathbf{z}_t^{t_0}\right) - \bar{Q}_N\left(t_0 \mid W, A=0, \mathbf{Z}_{t-1}, \mathbf{z}_t^{t_0}\right)\right) \right\}, \text{ for } t = 1, \ldots, t_0, \end{aligned}$$

and

$$\phi_{Z,NDE}(P)(t_0+1;W,\mathbf{Z}_{t_0}) \equiv \bar{Q}_N(t_0 \mid W, A=1, \mathbf{Z}_{t_0}) - \bar{Q}_N(t_0 \mid W, A=0, \mathbf{Z}_{t_0}).$$

Note that $\Psi_{NDE}(P) = E_{Q_W}(\phi_{Z,NDE}(P)(t=1;W))$, and the analog of the recursive relation in (11) applies to $\phi_{Z,NDE}(P)$.

On the other hand, for $t = 1, \ldots, t_0$, the *conditional natural indirect effect at t* is defined as

$$\phi_{Z,NIE}(P)(t; W, A = 1, \mathbf{Z}_{t-1}) - \phi_{Z,NIE}(P)(t; W, A = 0, \mathbf{Z}_{t-1}),$$

where

$$\begin{aligned}&\phi_{Z,NIE}(P)(t; W, A, \mathbf{Z}_{t-1}) \equiv \phi_{Z,1,A}(P)(t; W, \mathbf{Z}_{t-1})\\&= \sum_{\mathbf{z}_t^{t_0}} \prod_{t'=t}^{t_0} g_Z(z_{t'} \mid W, A, \mathbf{Z}_{t-1}, \mathbf{z}_t^{t'-1}, N_{t'-1} = 0) \bar{Q}_N\left(t_0 \mid W, A = 1, \mathbf{Z}_{t-1}, \mathbf{z}_t^{t_0}\right),\end{aligned}$$

and for $t = t_0 + 1$, we use the notation $\phi_{Z,NIE}(P)(t_0+1; W, A, \mathbf{Z}_{t_0}) \equiv \bar{Q}_N\left(t_0 \mid W, A = 1, \mathbf{Z}_{t_0}\right)$. It follows that $\Psi_{NIE}(P) = E_{Q_W}\left(\phi_{Z,NIE}(P)(t; W, A = 1) - \phi_{Z,NIE}(P)(t; W, A = 0)\right)$. Similar to the recursive relation in (11), $\phi_{Z,NIE}(P)$ satisfies $\phi_{Z,NIE}(P)(t-1; W, A, \mathbf{Z}_{t-2}) = E_{g_{Z,t-1}}\left(\phi_{Z,NIE}(P)(t; W, A, \mathbf{Z}_{t-1}) \mid W, A, \mathbf{Z}_{t-2}, N_{t-2} = 0\right)$

Theorem 2. *Let $\Psi_{a,a'} : \mathscr{M} \to \mathbb{R}$ be defined as in (10). Suppose the following are true for $P \in \mathscr{M}$:*

P1. There exists $0 < \delta_A < 1$ such that $g_A(A \mid W) > \delta_A$, a.e. over $\mathscr{A}$;

P2. There exists $0 < \delta_Z < 1$ such that $\inf_{t \in \{1,\ldots,t_0\}} g_Z(Z_t \mid W, A = a, \mathbf{Z}_{t-1}, N_{t-1} = 0) > \delta_Z$, a.e. over $\mathscr{Z}$;

P3. There exists $0 < \delta_N < 1$ such that $\inf_{t \in \{1,\ldots,t_0\}} 1 - Q_{dN}(t \mid W, A = a', \mathbf{Z}_t, N_{t-1} = 0) > \delta_N$.

The Efficient Influence Function of $\Psi_{a,a'}$ *at* P *is given by*

$$\begin{aligned}
&D^*_{a,a'}(P)(O) = -\sum_{t=1}^{t_0} I(N_{t-1}=0)\Bigg\{ \frac{I(A=a)}{g_A(a\mid W)} \prod_{t'=1}^{t} \frac{g_Z(Z_{t'}\mid W,A=a',\mathbf{Z}_{t'-1},N_{t'-1}=0)}{g_Z(Z_{t'}\mid W,A=a,\mathbf{Z}_{t'-1},N_{t'-1}=0)} \\
&\times \sum_{\mathbf{z}_{t+1}^{t_0}} \prod_{t'=t+1}^{t_0} g_Z(z_{t'}\mid W,A=a',\mathbf{Z}_t,\mathbf{z}_{t+1}^{t'-1},N_{t'-1}=0)\left(1-Q_{dN}(t'\mid W,A=a,\mathbf{Z}_t,\mathbf{z}_{t+1}^{t'})\right)\Bigg\} \\
&\times (dN_t - Q_{dN}(t\mid W,A=a,\mathbf{Z}_t)) \\
&+\sum_{t=1}^{t_0} I(N_{t-1}=0)\left\{\frac{I(A=a')}{g_A(a'\mid W)} \frac{\left(\phi_{Z,a,a'}(P)(t+1;W,\mathbf{Z}_t)-\phi_{Z,a,a'}(P)(t;W,\mathbf{Z}_{t-1})\right)}{\bar{Q}_N(t-1\mid W,A=a',\mathbf{Z}_{t-1})}\right\} \\
&+\phi_{Z,a,a'}(P)(t=1;W) - E_{Q_W}(\phi_{Z,a,a'}(P)(t=1;W)).
\end{aligned} \tag{14}$$

Proof. See appendix A2. □

The EIFs of both the natural direct (12) and indirect (13) effects can be derived from (14) by a simple application of the delta method. We state them in a corollary without proof.

Corollary 1. *Suppose the conditions P1 – P3 in theorem 2 hold for* $a,a' \in \{0,1\}$. *The efficient influence function of the natural direct effect (12) is given by*

$$\begin{aligned}
&D^*_{NDE}(P)(O) = D^*_{1,0}(P)(O) - D^*_{0,0}(P)(O) \\
&= -\sum_{t=1}^{t_0} I(N_{t-1}=0)\Bigg\{ \frac{2A-1}{g_A(A\mid W)} \prod_{t'=1}^{t} \frac{g_Z(Z_{t'}\mid W,A=0,\mathbf{Z}_{t'-1},N_{t'-1}=0)}{g_Z(Z_{t'}\mid W,A,\mathbf{Z}_{t'-1},N_{t'-1}=0)} \\
&\times \sum_{\mathbf{z}_{t+1}^{t_0}} \prod_{t'=t+1}^{t_0} g_Z(z_{t'}\mid W,A=0,\mathbf{Z}_t,\mathbf{z}_{t+1}^{t'-1},N_{t'-1}=0)\left(1-Q_{dN}(t'\mid W,A,\mathbf{Z}_t,\mathbf{z}_{t+1}^{t'})\right)\Bigg\} \\
&\times (dN_t - Q_{dN}(t\mid W,A,\mathbf{Z}_t)) \\
&+\sum_{t=1}^{t_0} I(N_{t-1}=0)\left\{\frac{I(A=0)}{g_A(0\mid W)} \frac{(\phi_{Z,NDE}(P)(t+1;W,\mathbf{Z}_t)-\phi_{Z,NDE}(P)(t;W,\mathbf{Z}_{t-1}))}{\bar{Q}_N(t-1\mid W,A=0,\mathbf{Z}_{t-1})}\right\} \\
&+\phi_{Z,NDE}(P)(t=1;W) - E_{Q_W}(\phi_{Z,NDE}(P)(t=1;W)),
\end{aligned} \tag{15}$$

and the efficient influence function of the natural indirect effect (13) is given by

$$
\begin{aligned}
& D^*_{NIE}(P)(O) = D^*_{1,1}(P)(O) - D^*_{1,0}(P)(O) \\
& = -\sum_{t=1}^{t_0} I(N_{t-1}=0)\,\frac{I(A=1)}{g_A(1\mid W)} \\
& \times \left\{ \sum_{\mathbf{z}_{t+1}^{t_0}} \left(\frac{G_Z\left(\mathbf{Z}_t,\mathbf{z}_{t+1}^{t_0}\mid W,A=1\right) - G_Z\left(\mathbf{Z}_t,\mathbf{z}_{t+1}^{t_0}\mid W,A=0\right)}{G_Z\left(\mathbf{Z}_t\mid W,A=1\right)} \prod_{t'=t+1}^{t_0} 1 - Q_{dN}(t'\mid W,A=1,\mathbf{Z}_t,\mathbf{z}_{t+1}^{t'}) \right) \right\} \\
& \times \left(dN_t - Q_{dN}(t\mid W,A=1,\mathbf{Z}_t)\right) \\
& + \sum_{t=1}^{t_0} I(N_{t-1}=0) \left\{ \frac{2A-1}{g_A(A\mid W)} \, \frac{(\phi_{Z,NIE}(P)(t+1;W,A,\mathbf{Z}_t) - \phi_{Z,NIE}(P)(t;W,A,\mathbf{Z}_{t-1}))}{\bar{Q}_N(t-1\mid W,A,\mathbf{Z}_{t-1})} \right\} \\
& + \left(\phi_{Z,NIE}(P)(t=1;W,A=1) - \phi_{Z,NIE}(P)(t=1;W,A=0)\right) \\
& - E_{Q_W}\left(\phi_{Z,NIE}(P)(t=1;W,A=1) - \phi_{Z,NIE}(P)(t=1;W,A=0)\right).
\end{aligned} \tag{16}
$$

The variances $Var_{P_0}(D^*_{a,a'}(P_0))$, $Var_{P_0}(D^*_{NDE}(P_0))$, and $Var_{P_0}(D^*_{NIE}(P_0))$ are generalized Cramer-Rao lower bounds for the asymptotic variances of the RAL estimators of $\Psi_{a,a'}(P_0)$, $\Psi_{NDE}(P_0)$, and $\Psi_{NIE}(P_0)$, respectively. Moreover, under our model, all RAL estimators will have their influence function given by the EIFs. Therefore, the asymptotic behavior of these estimators are governed by the EIFs.

2.2.2. Robustness of the efficient influence functions

In practice, it is often difficult to estimate the entire distribution P_0. In this section, we study the robustness properties of the EIFs (14), (15) and (16) against certain model mis-specifications. These will provide insight on the potential robustness that an estimator can offer. We begin with the robustness properties of $D^*_{a,a'}$.

Lemma 1. *Let $\Psi_{a,a'}(P)$ be as defined in (10); its efficient influence function under $\mathscr{M}$ is $D^*_{a,a'}(P)$, as given in (14).*

Suppose for $P_0 \in \mathscr{M}$, conditions P1 – P3 of theorem 2 hold. Then,

$$P_0 D^*_{a,a'}(Q,g,\Psi_{a,a'}(P_0)) = 0$$

if one of the following holds:

R1. $Q_{dN} = Q_{dN,0}$*, and* $\phi_{Z,a,a'}(P) = \phi_{Z,a,a'}(P_0)$.

R2. $Q_{dN} = Q_{dN,0}$*, and* $g_A = g_{A,0}$.

R3. $(g_A, g_Z) = (g_{A,0}, g_{Z,0})$ *and* $\phi_{Z,a,a'}(P) = \phi_{Z,a,a'}(g_{Z,0}, Q_{dN})$.

Proof. See Appendix A2. □

The robustness properties of D^*_{NDE} and D^*_{NIE} can be derived in an analogous manner as lemma 1, we will omit the proofs here and state the results in this corollary:

Corollary 2. *Let* $\Psi_{NDE}(P)$ *be as defined in (12) and* $\Psi_{NIE}(P)$ *be as defined in (13); their efficient influence functions under* $\mathscr{M}$ *are* $D^*_{NDE}(P)$ *and* $D^*_{NIE}(P)$*, as given by (15) and (16), respectively.*

Suppose for $P_0 \in \mathscr{M}$*, conditions P1 – P3 of theorem 2 hold for* $a, a' \in \{0, 1\}$*. Then,*

$$P_0 D^*_{NDE}(Q, g, \Psi_{NDE}(P_0)) = 0$$

if one of the following holds:

R1. $Q_{dN} = Q_{dN,0}$*, and* $\phi_{Z,NDE}(P) = \phi_{Z,NDE}(P_0)$*,*

R2. $Q_{dN} = Q_{dN,0}$*, and* $g_A = g_{A,0}$*,*

R3. $(g_A, g_Z) = (g_{A,0}, g_{Z,0})$ *and* $\phi_{Z,NDE}(P) = \phi_{Z,NDE}(g_{Z,0}, Q_{dN})$*;*

and

$$P_0 D^*_{NIE}(P, g, \Psi_{NIE}(P_0)) = 0$$

if one of the following holds:

R1. $Q_{dN} = Q_{dN,0}$*, and* $\phi_{Z,NIE}(P) = \phi_{Z,NIE}(P_0)$.

R2. $Q_{dN} = Q_{dN,0}$*, and* $g_A = g_{A,0}$*,*

R3. $(g_A, g_Z) = (g_{A,0}, g_{Z,0})$ *and* $\phi_{Z,NIE}(P) = \phi_{Z,NIE}(g_{Z,0}, Q_{dN})$.

Note that in deriving these robustness properties, conditions R1 in lemma 1 and corollary 2 do not require estimation of the mediator density per se, if one can estimate the true conditional functionals $\phi_{Z,-}(P_0)$ ($\phi_{Z,a,a'}(P_0)$, $\phi_{Z,NDE}(P_0)$, or $\phi_{Z,NIE}(P_0)$) using a regression-based estimator which, at each t, conditions on the history recorded thus far. Similarly, the derivation of conditions R2 only rely on the recursive property (11) of the conditional functionals $\phi_{Z,-}(P)$, which can be satisfied by a regression-based estimator. On the other hand, R3 does require $g_{Z,0}$ to be correct, and $\phi_{Z,-}(P)$ to yield the true conditional expectations over Z_t, even though its integrand $\bar{Q}_N$ may be mis-specified. These suggest that in applications where the mediator density is difficult to estimate, using a regression-based estimator (instead of a substitution-based estimator) of the conditional functionals $\phi_{Z,-}(P_0)$ may be a viable alternative.

Estimators which satisfy the EIF equations will also inherit their robustness properties. We will present four estimators in the next section, two of which are robust and locally efficient.

2.3 Estimators

In this section, we develop the g-computation, IPTW, A-IPTW and TMLE estimators for the natural direct effect (12); the corresponding estimators for the natural indirect effect (13) follow very similar steps — we summarize them in appendix A3. The g-computation and the IPTW (inverse probability-of-treatment weighted) estimators are consistent only if the estimates of all the relevant components of P_0 are consistent. On the other hand, the A-IPTW (augmented IPTW) and the TMLE (targeted maximum likelihood) estimators satisfy the efficient influence function equation, and hence remain unbiased under the model mis-specifications described in corollary 2. Under appropriate regularity conditions, they will be consistent and asymptotically linear (e.g. Bickel et al. (1997), van der Laan and Robins (2003), van der Laan and

Rose (2011)). If all the components needed for evaluation of $D^*(P_0)$ are consistently estimated, these estimators will also be efficient.

Let $\hat{Q}$ denote an estimating procedure for the component Q_0 of P_0, and $\hat{Q}_n \equiv \hat{Q}(P_n)$ denote the estimator that results from training $\hat{Q}$ on the empirical distribution P_n. Similar definitions apply to $\hat{g}$ and $\hat{g}_n$. Use the bar notation $\bar{\hat{Q}}_{N,n}(t \mid W,A,\mathbf{Z}_t)$ for the product $\prod_{t'=1}^{t} 1 - \hat{Q}_{dN,n}(t' \mid W,A,\mathbf{Z}_{t'},N_{t'-1} = 0)$, and use the notation $\hat{G}_{Z,n}(\mathbf{Z}_t \mid W,A)$ for $\prod_{t'=1}^{t} \hat{g}_{Z,n}(Z_{t'} \mid W,A,\mathbf{Z}_{t'-1},N_{t'-1} = 0)$. An estimator $\hat{\phi}_{Z,NDE,n}(\cdot)$ of $\phi_{Z,NDE}(g_{Z,0},\cdot)$ maps an estimator $\hat{Q}_{dN,n}$ of $Q_{dN,0}$ to an estimator $\hat{\phi}_{Z,NDE,n}(\hat{Q}_{dN,n})$ of $\phi_{Z,NDE}(g_{Z,0},Q_{dN,0})$. This estimating procedure can be substitution- or regression-based. For a substitution-based estimator, $\hat{\phi}_{Z,NDE,n}(\hat{Q}_{dN,n}) \equiv \phi_{Z,NDE}(\hat{g}_{Z,n},\hat{Q}_{dN,n})$. For a regression-based estimator, $\hat{\phi}_{Z,NDE,n}(\hat{Q}_{dN,n})(t;W,\mathbf{Z}_{t-1})$ regresses the difference

$$\left(\bar{\hat{Q}}_{N,n}(t_0 \mid W,A=1,\mathbf{z}_{t_0}) - \bar{\hat{Q}}_{N,n}(t_0 \mid W,A=0,\mathbf{z}_{t_0})\right)$$

on $(W,\mathbf{Z}_{t-1})$ among observations with $A=0$ that haven't failed by time $t-1$. When Z_t is high-dimensional, a regression-based $\hat{\phi}_{Z,NDE,n}$ may be easier to implement.

2.3.1. g-computation

Let $\hat{Q}_{dN,n}$ and $\hat{g}_{Z,n}$ be estimators of $Q_{dN,0}$ and $g_{Z,0}$, respectively. We can use these to obtain an estimate for $\Psi_{NDE}(P_0)$ by plugging them into (12):

$$\begin{aligned}&\hat{\Psi}^{gcomp}_{NDE}(P_n)\\ &\equiv \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{\mathbf{z}_{t_0}} \hat{G}_{Z,n}(\mathbf{z}_{t_0} \mid W_i,A=0)\left(\bar{\hat{Q}}_{N,n}(t_0 \mid W_i,A=1,\mathbf{z}_{t_0}) - \bar{\hat{Q}}_{N,n}(t_0 \mid W_i,A=0,\mathbf{z}_{t_0})\right)\right).\end{aligned}$$

More generally, given estimators $\hat{Q}_{dN,n}$ and $\hat{\phi}_{Z,NDE,n}$, a g-computation estimate of $\Psi_{NDE}(P_0)$ can be obtained by

$$\hat{\Psi}^{gcomp}_{NDE}(P_n) \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{\phi}_{Z,NDE,n}(\hat{Q}_{dN,n})(t=1;W_i).$$

The consistency of the g-computation estimator relies on consistent estimation of $Q_{dN,0}$ and $\phi_{Z,NDE}(P_0)$.

Compared to the IPTW estimator below, in the presence of near positivity violation, the g-computation estimator remains bounded within the range of Ψ_{NDE}. Nonetheless, lack of experimental support can still manifest in poor estimates of Q_{dN} and $\phi_{Z,NDE}$.

2.3.2. IPTW

Instead of estimating the failure probability $Q_{dN,0}$, one may wish to employ the researcher's knowledge about the treatment assignment. Consider the following function:

$$D_{NDE,IPTW}(P) = \frac{2A-1}{g_A(A \mid W)} \prod_{t=1}^{t_0} \frac{g_Z(Z_t \mid W, A=0, \mathbf{Z}_{t-1}, N_{t-1})}{g_Z(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1})} I(N_{t_0}=0).$$

Notice that $\Psi_{NDE}(P_0) = P_0 D_{NDE,IPTW}(P_0)$:

$$\begin{aligned}
&P_0 D_{NDE,IPTW}(P_0) = E_{P_0}\left(\frac{2A-1}{g_{A,0}(A \mid W)} \prod_{t=1}^{t_0} \frac{g_{Z,0}(Z_t \mid W, A=0, \mathbf{Z}_{t-1}, N_{t-1})}{g_{Z,0}(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1})} I(N_t=0) \right) \\
&= E_{Q_W}\left(\sum_{\mathbf{z}_{t_0}} \prod_{t=1}^{t_0} g_{Z,0}(z_t \mid W, A=1, \mathbf{z}_{t-1}, N_{t-1}=0) \frac{g_{Z,0}(z_t \mid W, A=0, \mathbf{z}_{t-1}, N_{t-1}=0)}{g_{Z,0}(z_t \mid W, A=1, \mathbf{z}_{t-1}, N_{t-1}=0)} \right. \\
&\left. \times (1 - Q_{dN,0}(t \mid W, A=1, \mathbf{z}_t, N_{t-1}=0)) \right) \\
&- E_{Q_W}\left(\sum_{\mathbf{z}_{t_0}} \prod_{t=1}^{t_0} g_{Z,0}(z_t \mid W, A=0, \mathbf{z}_{t-1}, N_{t-1}=0) \frac{g_{Z,0}(z_t \mid W, A=0, \mathbf{z}_{t-1}, N_{t-1}=0)}{g_{Z,0}(z_t \mid W, A=0, \mathbf{z}_{t-1}, N_{t-1}=0)} \right. \\
&\left. \times (1 - Q_{dN,0}(t \mid W, A=0, \mathbf{z}_t, N_{t-1}=0)) \right) \\
&= \Psi_{NDE}(P_0)
\end{aligned}$$

Therefore, given estimators $\hat{g}_{A,n}$ and $\hat{g}_{Z,n}$ of $g_{A,0}$ and $g_{Z,0}$, respectively, an estimator of $\Psi_{NDE}(P_0)$ can be obtained using:

$$\hat{\Psi}_{NDE}^{IPTW}(P_n) \equiv \frac{1}{n} \sum_{i=1}^{n} \frac{2A_i - 1}{\hat{g}_{A,n}(A_i \mid W_i)} \prod_{t=1}^{t_0} \frac{\hat{g}_{Z,n}(Z_{i,t} \mid W_i, A=0, \mathbf{Z}_{i,t-1}, N_{i,t-1})}{\hat{g}_{Z,n}(Z_{i,t} \mid W_i, A_i, \mathbf{Z}_{i,t-1}, N_{i,t-1})} I(N_{i,t_0}=0). \quad (17)$$

As noted earlier, if $N_{i,t-1} \neq 0$, $\hat{g}_{Z,n}(Z_{i,t} \mid W_i, A, \mathbf{Z}_{i,t-1}, N_{i,t-1})$ takes the value of 1, since $Z_{i,t}$ would be deterministically assigned a degenerate value. This way, (17) is well-defined. This estimator is consistent if $g_0 = (g_{Z,0}, g_{A,0})$ is consistently estimated.

Compared to the g-computation estimator, the IPTW is more sensitive to near positivity violations. In particular, the parameter estimate is unsheltered from the impact of small denominator values. One way to reduce the variance of the resulting estimate is to truncate the values of the denominators. This option comes at the expense that the truncated denominators are biased estimates of g_0. Bembom and van der Laan (2008) propose a data-adaptive selection of the truncation level to obtain an optimal finite bias-variance tradeoff for the parameter of interest.

2.3.3. A-IPTW

To gain protection against certain model mis-specifications, one can exploit the robustness properties of the efficient influence functions. In particular, the EIF $D^*(P)$ of a parameter $\Psi(P)$ can be used as an estimating function of $\Psi(P)$ (e.g. Robins (1999), Robins and Rotnitzky (2001), van der Laan and Robins (2003)), if (i) $D^*(P)$ can be expressed as a function of Ψ and some nuisance parameter η, i.e. $D^*(P) = D(\Psi(P), \eta(P))$, for some function D, and (ii) the solution to the resulting equation in the variable Ψ is unique. When these requirements hold, an estimate $\hat{\Psi}$ of the parameter is defined as the solution of the resulting estimating equation $P_n D^*(\hat{\eta}(P_n), \hat{\Psi}) = 0$. This $\hat{\Psi}$ is also known to as the A-IPTW estimator.

For the natural direct effect parameter in (12), the efficient influence function D^*_{NDE} in (15) is a well-defined estimating function with nuisance parameters $Q(P)$, $g(P)$, and $\phi_{Z,NDE}(P)$. Let $\hat{Q}_{dN,n}$, $\hat{g}_{Z,n}$, $\hat{g}_{A,n}$ and $\hat{\phi}_{Z,NDE,n}$ be their respective estimators.

The A-IPTW estimator is given by

$$
\begin{aligned}
&\hat{\Psi}_{NDE}^{AIPTW}(P_n)\\
&= \frac{1}{n}\sum_{i=1}^{n}\Bigg\{-\sum_{t=1}^{t_0} I(N_{i,t-1}=0)\Bigg(\frac{(2A_i-1)}{\hat{g}_{A,n}(A_i \mid W_i)}\prod_{t'=1}^{t}\frac{\hat{g}_{Z,n}(Z_{i,t'} \mid W_i, A=0, \mathbf{Z}_{i,t'-1}, N_{t'-1}=0)}{\hat{g}_{Z,n}(Z_{i,t'} \mid W_i, A_i, \mathbf{Z}_{i,t'-1}, N_{t'-1}=0)}\\
&\times \sum_{\mathbf{z}_{t+1}^{t_0}}\prod_{t'=t+1}^{t_0}\hat{g}_{Z,n}(Z_{i,t'} \mid W_i, A=0, \mathbf{Z}_{i,t-1}, \mathbf{z}_{t+1}^{t'-1}, N_{t'-1}=0)\left(1-\hat{Q}_{dN,n}(t' \mid W_i, A_i, \mathbf{Z}_{i,t-1}, \mathbf{z}_{t+1}^{t'})\right)\Bigg)\\
&\times \left(dN_{i,t-1}-\hat{Q}_{dN,n}(t \mid W_i, A_i, \mathbf{Z}_{i,t-1})\right)\\
&+\sum_{t=1}^{t_0} I(N_{i,t-1}=0)\frac{I(A_i=0)}{\hat{g}_{A,n}(0 \mid W_i)}\frac{\left(\hat{\phi}_{Z,NDE,n}(\hat{Q}_{dN,n})(t+1; W_i, \mathbf{Z}_{i,t-1})-\hat{\phi}_{Z,NDE,n}(\hat{Q}_{dN,n})(t; W_i, \mathbf{Z}_{i,t-1})\right)}{\bar{\hat{Q}}_{N,n}(t-1 \mid W_i, A_i=0, \mathbf{Z}_{i,t-1})}\\
&+\hat{\phi}_{Z,NDE,n}(\hat{Q}_{dN,n})(t=1; W_i)\Bigg\}.
\end{aligned}
$$

This estimator is multiply robust in the sense that if either one of the conditions R1, R2, or R3 in corollary 2 hold at the limit of these likelihood estimates, then $\hat{\Psi}_{NDE}^{AIPTW}(P_n)$ is an asymptotically unbiased estimator of $\Psi_{NDE}(P_0)$. Therefore, it offers more protection against model mis-specifications than the g-computation and IPTW estimators.

When near positivity violations are present, the remarks given to the IPTW estimator equally apply here: parameters estimates are unguarded against the impact of small denominator values; truncation of the denominators can reduce variance, but the truncation level should be selected to optimize the bias-variance tradeoff for the parameter of interest.

2.3.4. TMLE

To maximize finite sample gain and provide more stable estimates in the presence of near positivity violations, one can make use of the substitution principle. The targeted maximum likelihood estimation (TMLE, van der Laan and Rubin (2006)) provides a substitution-based estimator which also satisfies the EIF equation, thereby remaining unbiased under model mis-specifications. Under this framework, for each relevant component P_j of P, one defines a uniformly bounded (w.r.t. the supremum norm) *loss function* L_j satisfying $P_{j,0} = \arg\min_{P_j \in \mathscr{P}_j} P_0 L_j(P_j)$, and a one-dimensional *parametric working submodel* $\{P_j(\varepsilon_j) : \varepsilon_j\} \subset \mathscr{M}$, passing through P_j at $\varepsilon_j = 0$, with score $D^*_j(P)$ at $\varepsilon_j = 0$ that satisfies $\langle \frac{d}{d\varepsilon_j} L_j\left(P_j(\varepsilon_j)\right) |_{\varepsilon_j=0}\rangle \supset \langle D^*_j(P)\rangle$, where $\langle h\rangle$ denotes the linear span of a vector h. These result in a least favorable parametric submodel $P(\varepsilon)$ through P. For given initial estimator $\hat{P}_n$ of P_0, the parameter ε is fitted to minimize the empirical risk of $\hat{P}_n(\varepsilon)$, producing an updated estimator $\hat{P}_n(\hat{\varepsilon})$ of P_0. This updating process is repeated until $\hat{\varepsilon} \approx 0$. The final updated estimator $\hat{P}^*_n$ of P_0 is then used to obtain a substitution estimator $\Psi(\hat{P}^*_n)$ of $\Psi(Q_0)$. By its construction, the estimator $\hat{P}^*_n$ satisfies the efficient influence function equation $P_n D^*(\hat{P}^*_n) = 0$.

To construct the TMLE estimator for the natural direct effect (12), we first decompose the EIF (15) into its orthogonal components:

$$
\begin{aligned}
& D^*_{NDE,N}(P) \equiv \sum_{t=1}^{t_0} D^*_{NDE,dN_t}(P) \\
& = -\sum_{t=1}^{t_0} I(N_{t-1}=0) \left\{ \frac{2A-1}{g_A(A \mid W)} \prod_{t'=1}^{t} \frac{g_Z(Z_{t'} \mid W, A=0, \mathbf{Z}_{t'-1}, N_{t'-1}=0)}{g_Z(Z_{t'} \mid W, A, \mathbf{Z}_{t'-1}, N_{t'-1}=0)} \right. \\
& \left. \times \sum_{\mathbf{z}_{t+1}^{t_0}} \prod_{t'=t+1}^{t_0} g_Z(z_{t'} \mid W, A=0, \mathbf{Z}_t, \mathbf{z}_{t+1}^{t'-1}, N_{t'-1}=0) \left(1 - Q_{dN}(t' \mid W, A, \mathbf{Z}_t, \mathbf{z}_{t+1}^{t'})\right) \right\} \\
& \times \left(dN_t - Q_{dN}(t \mid W, A, \mathbf{Z}_t, N_{t-1}=0)\right), \\
& D^*_{NDE,Z}(P) \equiv \sum_{t=1}^{t_0} D^*_{NDE,Z_t}(P) \\
& = \sum_{t=1}^{t_0} I(N_{t-1}=0) \left\{ \frac{I(A=0)}{g_A(0 \mid W)} \frac{\left(\phi_{Z,NDE}(P)(t+1; W, \mathbf{Z}_t) - \phi_{Z,NDE}(P)(t; W, \mathbf{Z}_{t-1})\right)}{\bar{Q}_N(t-1 \mid W, A=0, \mathbf{Z}_{t-1})} \right\}, \\
& D^*_{NDE,W}(P) \equiv \phi_{Z,NDE}(P)(t=1; W) - E_{Q_W} \phi_{Z,NDE}(P)(t=1; W).
\end{aligned}
$$

Note that the empirical marginal distribution $\hat{Q}_{W,n}$ of W is a consistent estimator of $Q_{W,0}$ that readily satisfies the equation $P_n D^*_{NDE,W}(\phi_{Z,NDE}(P), \hat{Q}_{W,n}) = 0$ for any $\phi_{Z,NDE}(P)$. Hence, the proposed estimator will focus on targeted estimation of $Q_{dN,0}$

and $\phi_{Z,NDE}(P_0)$.

To simplify notation, we suppress $Q_{dN}(t \mid W, A, \mathbf{Z}_t, N_{t-1} = 0)$ into $Q_{dN}(t)$ for the remaining of this section, it should be understood that $Q_{dN}(t)$ is always a function of the parents of dN_t. For every $t = 1, \ldots, t_0$, let the loss function for $Q_{dN}(t)$ be the minus-loglikelihood:

$$L_{dN_t}(Q_{dN}(t))(O) = I(N_{t-1} = 0) \log \left(Q_{dN}(t)^{dN_t} (1 - Q_{dN}(t))^{1-dN_t} \right).$$

Under this loss function, consider the logistic working submodel

$$Q_{dN}(t)(\varepsilon) \equiv expit \left(logit \left(Q_{dN}(t) \right) + \varepsilon C_{dN}(g, Q_{dN})(t) \right),$$

where

$$\begin{aligned} & C_{dN}(g, Q_{dN})(t) \\ & \equiv \frac{2A-1}{g_A(A \mid W)} \prod_{t'=1}^{t} \frac{g_Z(Z_{t'} \mid W, A = 0, \mathbf{Z}_{t'-1}, N_{t'-1} = 0)}{g_Z(Z_{t'} \mid W, A, \mathbf{Z}_{t'-1}, N_{t'-1} = 0)} \\ & \times \sum_{\mathbf{z}_{t+1}^{t_0}} \prod_{t'=t+1}^{t_0} g_Z(z_{t'} \mid W, A = 0, \mathbf{Z}_t, \mathbf{z}_{t+1}^{t'-1}, N_{t'-1} = 0) \left(1 - Q_{dN}(t' \mid W, A, \mathbf{Z}_t, \mathbf{z}_{t+1}^{t'}, N_{t'-1} = 0) \right) \end{aligned} \tag{18}$$

It is suppressed in this notation that $C_{dN}(g, Q_{dN})(t)$ is a function of $(W, A, \mathbf{Z}_t)$. Note that for a given t, $C_{dN}(g, Q_{dN})(t)$ only depends on $Q_{dN}(j)$ for $j > t$.

After a fixed linear transformation, we may assume that the difference

$$\phi_{Z,NDE}(P)(t_0 + 1; W, \mathbf{Z}_{t_0}) \equiv \bar{Q}_N(t_0 \mid W, 1, \mathbf{Z}_{t_0}) - \bar{Q}_N(t_0 \mid W, 0, \mathbf{Z}_{t_0})$$

is bounded in the unit interval. Recall that the conditional natural direct effects satisfy the recursive relation

$$\phi_{Z,NDE}(P)(t; W, \mathbf{Z}_{t-1}) = E_{g_{Z,t}} \left(\phi_{Z,NDE}(P)(t+1; W, \mathbf{Z}_t) \mid W, A = 0, \mathbf{Z}_{t-1}, N_{t-1} = 0 \right).$$

We suppress the notation $\phi_{Z,NDE}(P)(t; W, \mathbf{Z}_{t-1})$ into $\phi_{Z,NDE}(P)(t)$. Consider the following loss function for $\phi_{Z,NDE}(P)(t)$ at $t = 1, \ldots, t_0$:

$$\begin{aligned} & L_{Z_t}\left(\phi_{Z,NDE}(P)(t) \right) \\ & \equiv -I(A = 0) I(N_{t-1} = 0) \log \left(\left(\phi_{Z,NDE}(P)(t) \right)^{\phi_{Z,NDE}(P)(t+1)} \left(1 - \phi_{Z,NDE}(P)(t) \right)^{1-\phi_{Z,NDE}(P)(t+1)} \right), \end{aligned}$$

with parametric working submodel given by

$$\phi_{Z,NDE}(P)(t)(\varepsilon) = expit\left(logit\left(\phi_{Z,NDE}(P)(t)\right) + \varepsilon C_Z(g,Q_{dN})(t)\right),$$

where

$$C_Z(g,Q_{dN})(t) = \frac{1}{g_A(0|W)\bar{Q}_N(t-1 \mid W,A=0,\mathbf{Z}_{t-1})}. \tag{19}$$

Implementation

Let $\hat{g}_{A,n}$, $\hat{g}_{Z,n}$ and $\hat{Q}_{dN,n}$ be initial estimators of g_0, $g_{Z,0}$ and $Q_{dN,0}$, respectively.

1. Starting with $t=t_0$, $C_{dN}(\hat{g}_n,\hat{Q}_{dN,n})(t_0) \equiv \frac{(2A-1)}{\hat{g}_{A,n}(A|W)}\prod_{t'=1}^{t_0}\frac{\hat{g}_{Z,n}(Z_{t'}|W,A=0,\mathbf{Z}_{t'-1},N_{t'-1}=0)}{\hat{g}_{Z,n}(Z_{t'}|W,A,\mathbf{Z}_{t'-1},N_{t'-1}=0)}$ and an optimal ε for $\hat{Q}_{dN,n}(t_0)$ is given by $\hat{\varepsilon}^*_{dN,t_0} = \arg\min_\varepsilon P_n L_{dN_{t_0}}\left(\hat{Q}_{dN,n}(t_0)(\varepsilon)\right)$. This can be obtained using standard software by performing a generalized linear regression $dN_{t_0} \sim offset(\hat{Q}_{dN,n}(t_0)) + C_{dN}(\hat{g}_n,\hat{Q}_{dN,n})(t_0)$ with logit link, where $offset(h)$ specifies to the program that the term h is an intercept in the model, on the subpopulation with $N(t_0-1)=0$. The fitted coefficient for the C_{dN} term in the model fit is the optimal $\hat{\varepsilon}^*_{dN,t_0}$. This provides an TMLE estimate $\hat{Q}^*_{dN,n}(t_0) \equiv \hat{Q}_{dN,n}(t_0)(\hat{\varepsilon}^*_{dN,t_0})$ for $Q_{dN,0}(t_0)$.

2. For each $1 \le t < t_0$, let $\hat{Q}^*_{dN,n}$ denote the vector of TMLE estimates

$$\hat{Q}^*_{dN,n} \equiv (\hat{Q}^*_{dN,n}(t_0), \hat{Q}^*_{dN,n}(t_0-1), \ldots, \hat{Q}^*_{dN,n}(t+1))$$

obtained thus far. We use these to construct $C_{dN}(\hat{g}_n,\hat{Q}^*_{dN,n})(t)$ as prescribed in (18). The optimal ε for $\hat{Q}_{dN,n}(t)$ is given by $\hat{\varepsilon}^*_{dN,t} = \arg\min_\varepsilon P_n L_{dN_t}\left(\hat{Q}_{dN,n}(t)(\varepsilon)\right)$.

The step 2 above updates $\hat{Q}_{dN,n}(t)$ sequentially in the order of descending t. Once we have obtained all the t_0 updates, let $\hat{Q}^*_{dN,n}$ to represent the TMLE estimator for the function $Q_{dN,0}$ at times $t=1,\ldots,t_0$. The same bar notation applies to this estimator: $\bar{\hat{Q}}^*_{N,n}(t \mid W,A,\mathbf{Z}_{t-1}) \equiv \prod_{t'=1}^{t} 1-\hat{Q}^*_{dN,n}(t' \mid W,A,\mathbf{Z}_{t'-1},N_{t'-1}=0)$.

Let $\hat{\phi}_{Z,NDE,n}(\cdot)$ be an estimating procedure for $\phi_{Z,NDE}(P_0)$ which maps the TMLE estimator $\hat{Q}^*_{dN,n}$ to an initial estimator $\hat{\phi}_{Z,NDE,n}(\hat{Q}^*_{dN,n})$ of $\phi_{Z,NDE}(g_{Z,0}, Q_{dN,0})$. The next steps will update $\hat{\phi}_{Z,NDE,n}(\hat{Q}^*_{dN,n})$ towards optimal bias-variance tradeoff for the parameter of interest. We suppress the notation $\hat{\phi}_{Z,NDE,n}(\hat{Q}^*_{dN,n})(t; W, \mathbf{Z}_{t-1})$ into $\hat{\phi}_{Z,NDE,n}(t)$, it should be understood that at this stage we always use the updated TMLE estimator $\hat{Q}^*_{dN,n}$ wherever Q_{dN} is involved.

3. Define

$$\hat{\phi}^*_{Z,NDE,n}(t_0+1; W, \mathbf{Z}_{t_0}) \equiv \hat{\bar{Q}}^*_{N,n}(t_0 \mid W, A=1, \mathbf{Z}_{t_0}) - \hat{\bar{Q}}^*_{N,n}(t_0 \mid W, A=0, \mathbf{Z}_{t_0})$$

   After a proper linear transformation, we may assume that $\hat{\phi}^*_{Z,NDE,n}(t_0+1; W, \mathbf{Z}_{t_0})$ is bounded in the unit interval.

4. For each $1 \leq t \leq t_0$, suppose we have obtained the TMLE estimator $\hat{\phi}^*_{Z,NDE,n}(t+1)$ for $\phi_{Z,NDE}(P_0)(t+1)$. The parametric submodel $\hat{\phi}_{Z,NDE}(t)(\varepsilon)$ is constructed using $C_Z(\hat{g}_n, \hat{Q}^*_{dN,n})(t)$ prescribed in (19), and the optimal ε is given by

$$\hat{\varepsilon}^*_{Z_t} = \arg\min_{\varepsilon} P_n L_{Z_t}\left(\hat{\phi}_{Z,NDE,n}(t)(\varepsilon)\right).$$

   This can be obtained using standard software by performing a generalized linear regression $\hat{\phi}^*_{Z,NDE,n}(t+1) \sim offset(\hat{\phi}_{Z,NDE,n}(t)) + C_Z(\hat{g}_n, \hat{Q}^*_{dN,n})(t)$ with logit link, on the subpopulation with $N_{t-1}=0$ and $A=0$.

   This yields the TMLE estimator $\hat{\phi}^*_{Z,NDE,n}(t) \equiv \hat{\phi}_{Z,NDE,n}(t)(\hat{\varepsilon}^*_{Z_t})$ of $\phi_{Z,NDE}(P_0)(t)$.

5. We perform the updates in step 4 sequentially in order of decreasing t. Once, we have obtained the TMLE estimator $\hat{\phi}^*_{Z,NDE,n}(t=1; W)$ (and applied the necessary inverse linear transformation), the TMLE estimate of the natural direct effect is given by

$$\hat{\Psi}^{TMLE}_{NDE}(P_n) \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{\phi}^*_{Z,NDE,n}(t=1; W_i).$$

Together with the initial estimates $\hat{g}_n$, the updated components $(\hat{Q}^*_{dN,n}, \hat{\phi}^*_{Z,NDE,n})$ satisfy $P_n D^*_{NDE}(\hat{g}_n, \hat{Q}^*_{dN,n}, \hat{\phi}^*_{Z,NDE,n}) = 0$. Therefore, the estimator $\hat{\Psi}^{TMLE}_{NDE}(P_n)$ is multiply robust in the sense that if either one of the conditions R1, R2 or R3 of corollary 2 hold at the limit of these likelihood estimates, then it is asymptotically unbiased.

In terms of finite sample performance, the logistic parametric working submodels and the substitution principle ensure that the resulting estimates are within proper bounds even in the presence of small denominator values in the EIF. This aims to provide some finite sample gain in the presence of near positivity violations. However, lack of experimental support can still manifest in poor initial estimates of the likelihood and poor fits for ε.

2.3.5. Simulations

In this section, we evaluate with simulations the performance of these four estimators under the three types of model mis-specifications in corollary 2. We expect to see A-IPTW and TMLE provide bias reduction over a mis-specified g-computation or IPTW estimators.

Consider the following data generating distribution

$W_1 \sim U(0,2)$;

$W_2 \sim Bern(0.5)$;

$A \sim Bern\left(expit(1 - W_1 - 0.5W_2)\right)$;

Conditional on $(W, A, \mathbf{Z}_{t-1}, N_{t-1} = 0)$, $Z_t \in \{0,1,2\} \sim$

$$Multinom\begin{pmatrix} p(Z_t = 0) = expit(0.5 - A - 2W_1 + 0.6W_2 + 0.2\sum_{t'=1}^{t-1} Z_{t'}), \\ p(Z_t = 1 | Z_t \neq 0) = expit(0.5 - 0.8A - 2W_1 + 0.8W_2 + 0.1\sum_{t'=1}^{t-1} Z_{t'}), \end{pmatrix},$$

and

$$(dN_t \mid W, A, \mathbf{Z}_t, N_{t-1} = 0) \sim Bern\Big(expit(0.3t - 3A - 3W_1 + 0.8W_2 + 0.2\sum_{t'=0}^{t} Z_{t'})\Big),$$

where $\sum_{t'=1}^{t-1} Z_{t'} \equiv 0$ for $t = 1$.

The survival threshold of interest is $t_0 = 3$. The parameter of interest $\Psi_{NDE}(P_0)$ for $t_0 = 3$ has value 0.4196206, the variance of its EIC is $Var_{P_0}(D^*) = 0.5288578$.

We consider the following model mis-specifications:

- The mis-specified model for g_A only adjusts for W_2.
- The mis-specified model for $Q_{dN}(t \mid W, A, \mathbf{Z}_t, N_{t-1} = 0)$ only adjusts for A, W_2 and Z_t.
- The mis-specified model for $g_Z(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1} = 0)$ only adjusts for A and Z_{t-1}.

For each of the model specifications in corollary 2, we obtain initial estimates $\hat{g}_{A,n}$, $\hat{g}_{Z,n}$ and $\hat{Q}_{dN,n}$. A plug-in procedure $\hat{\phi}_{Z,NDE,n}(\cdot) \equiv \phi_{Z,NDE}(\hat{g}_{Z,n}, \cdot)$ is used to estimate $\phi_{Z,NDE}(P_0)$.

Results

For each sample size $n = 500, 2000$, we generated 2000 datasets. Bias, variance and mse for each sample size are estimated over the 2000 datasets. In the table 2 below, legend for model specifications are as follows:

Table 1: Model specification

notation	model specifications
qyc.gzc.gac	correct Q_{dN}, correct g_Z, correct g_A
qyc.gzc.gam	correct Q_{dN}, correct g_Z, mis-specified g_A
qyc.gzm.gac	correct Q_{dN}, mis-specified g_Z, correct g_A
qym.gzc.gac	mis-specified Q_{dN}, correct g_Z, correct g_A

As predicted by theory, both A-IPTW and TMLE provide bias reduction over mis-specified g-computation estimators (qyc.gzm.gac and qym.gzc.gac) and mis-specified IPTW estimators (qyc.gzc.gam and qyc.gzm.gac). In cases where the g-computation and IPTW estimators are correctly specified (qyc.gzc. and gzc.gac., respectively), using A-IPTW and TMLE with a mis-specified third component still gives estimates very close to the truth. Therefore, without any knowledge on the consistency of the initial estimates of the likelihood components, applying a robust procedure (A-IPTW or TMLE) onto these initial estimates would provide protection against certain types of misspecification. Note also that when all relevant components of P_0 are correctly specified (qyc.gzc.gac), the sample variances of TMLE and A-IPTW are close to the semiparametric efficiency bounds after scaling by sample size.

Table 2: Sample sizes n=500, 2000. Value of parameter is $\psi_0 = 0.4196206$. $var(D^*)/500 = 1.058 \times 10^{-3}$, and $var(D^*)/2000 = 2.644 \times 10^{-4}$.

	$\hat{\psi}_n$		Bias		Var		M.S.E.	
n	500	2000	500	2000	500	2000	500	2000
qyc.gzc.gac								
gcomp	4.199e-01	4.194e-01	2.722e-04	-2.478e-04	5.150e-04	1.242e-04	5.151e-04	1.243e-04
iptw	4.197e-01	4.195e-01	1.132e-04	-1.289e-04	3.209e-03	7.359e-04	3.209e-03	7.359e-04
a-iptw	4.202e-01	4.193e-01	6.041e-04	-3.239e-04	1.091e-03	2.722e-04	1.092e-03	2.723e-04
tmle	4.198e-01	4.193e-01	1.637e-04	-3.395e-04	1.116e-03	2.719e-04	1.116e-03	2.720e-04
qyc.gzc.gam								
gcomp	4.199e-01	4.194e-01	2.722e-04	-2.478e-04	5.150e-04	1.242e-04	5.151e-04	1.243e-04
iptw	3.136e-01	3.132e-01	-1.060e-01	-1.064e-01	3.472e-03	8.459e-04	1.471e-02	1.216e-02
a-iptw	4.202e-01	4.194e-01	5.399e-04	-2.681e-04	1.279e-03	3.189e-04	1.279e-03	3.189e-04
tmle	4.205e-01	4.195e-01	8.382e-04	-1.382e-04	1.089e-03	2.700e-04	1.089e-03	2.700e-04
qyc.gzm.gac								
gcomp	4.065e-01	4.061e-01	-1.310e-02	-1.350e-02	4.997e-04	1.187e-04	6.712e-04	3.011e-04
iptw	4.174e-01	4.138e-01	-2.270e-03	-5.845e-03	2.364e-03	6.061e-04	2.369e-03	6.403e-04
a-iptw	4.201e-01	4.192e-01	4.897e-04	-4.683e-04	9.467e-04	2.409e-04	9.469e-04	2.411e-04
tmle	4.195e-01	4.190e-01	-1.506e-04	-6.192e-04	9.063e-04	2.320e-04	9.063e-04	2.324e-04
qym.gzc.gac								
gcomp	2.851e-01	2.850e-01	-1.345e-01	-1.346e-01	9.253e-04	2.296e-04	1.902e-02	1.834e-02
iptw	4.197e-01	4.195e-01	1.132e-04	-1.289e-04	3.209e-03	7.359e-04	3.209e-03	7.359e-04
a-iptw	4.200e-01	4.194e-01	3.978e-04	-2.459e-04	1.176e-03	2.837e-04	1.176e-03	2.838e-04
tmle	4.199e-01	4.193e-01	2.589e-04	-3.384e-04	1.114e-03	2.751e-04	1.114e-03	2.753e-04

3. RIGHT CENSORING

Up to now, we have assumed that there is no right censoring and all failure times are observed. In this section, we consider the situation where right censoring is present (e.g. lost-to-follow up, or study ended before the event occurs). Ideally, one would like to observe all the failure times; therefore, regardless of the treatment levels and mediator distributions, the interventions of interest always disallow censoring. By regarding censoring as a intervention variables, the same concepts in section 2 can be applied here.

Let C denote the first visit where an individual is right censored. We refer to this as the *censoring time*. Let $\tilde{T} \equiv min(T,C)$ be an individual's last observed visit, and $\Delta \equiv I(T \leq C)$ be the indicator that the failure time was observed (i.e. the subject was not censored). The observed data structure now consists of $O = (W,A,Z_1,\ldots,Z_{\tilde{T}},\tilde{T},\Delta)$. Let $N_t \equiv I(\tilde{T} \leq t, \Delta = 1)$ and $A_{C,t} \equiv I(\tilde{T} \leq t, \Delta = 0)$ denote two processes that jump to 1 at observed failure time and censoring time, respectively. Let $dN_t \equiv I(\tilde{T} = t, \Delta = 1)$ and $dA_{C,t} \equiv I(\tilde{T} = t, \Delta = 0)$ be the event and censoring indicators, respectively. The data structure can be represented as $O = (W,A,(Z_t,dN_t,dA_{C,t} : t = 1,\ldots,\tau))$, where for $t > \tilde{T}$, Z_t is given a degenerate value that is outside of $\mathscr{Z}$. Let P_0 denote the distribution of O. The data consists of n i.i.d. copies of O.

Let the following NPSEM encode the time-ordering of the variables:

$$\begin{aligned}
W &= f_W(U_W) \\
A &= f_A(W,U_A) \\
Z_t &= f_{Z_t}(W,A,\mathbf{Z}_{t-1},N_{t-1},A_{C,t-1},U_{Z_t}), \text{ for } t = 1,\ldots,\tau \\
dN_t &= f_{dN_t}(W,A,\mathbf{Z}_t,N_{t-1},A_{C,t-1},U_{dN_t}), \text{ for } t = 1,\ldots,\tau \\
dA_{C,t} &= f_{dA_{C,t}}(W,A,\mathbf{Z}_t,N_t,A_{C,t-1},U_{dA_{C,t}}), \text{ for } t = 1,\ldots,\tau. \qquad (20)
\end{aligned}$$

To save space, we may sometimes write $\tilde{T} > t$ in place of $(N_t = 0, A_{C,t} = 0)$.

According to the time-ordering, the likelihood of the observed data distribution P_0 decomposes into

$$\begin{aligned}
p_0(O) &= p_0(W)p_0(A \mid W) \\
&\times \prod_{t=1}^{\tau} \Big(p_0(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1}, A_{C,t-1}) p_0(dN_t \mid W, A, \mathbf{Z}_t, N_{t-1}, A_{C,t-1}) \\
&\times p_0(dA_{C,t} \mid W, A, \mathbf{Z}_t, N_t, A_{C,t-1}) \Big).
\end{aligned}$$

Let $g_{C,0}(t \mid W, A, \mathbf{Z}_t, N_t, A_{C,t-1}) \equiv p_0(dA_{C,t} = 1 \mid W, A, \mathbf{Z}_t, N_t, A_{C,t-1})$. We also modify previous notations to incorporate censoring: $g_{Z,0}(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1}, A_{C,t-1}) \equiv p_0(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1}, A_{C,t-1})$, $g_0 \equiv (g_{A,0}, g_{C,0}, g_{Z,0})$, and $Q_{dN,0}(t \mid W, A, \mathbf{Z}_t, N_{t-1}, A_{C,t-1}) \equiv p_0(dN_t = 1 \mid W, A, \mathbf{Z}_t, N_{t-1}, A_{C,t-1})$.

Let $Z_t(a', \Delta = 1)$ and $dN_t(a', \Delta = 1)$ denote the counterfactual mediator and event indicator under an intervention which sets $A = a'$ and $\mathbf{dA_C = 0}$. Let $g_{Z(a',\Delta=1)}$ denote the conditional distribution of $Z_t(a', \Delta = 1)$, i.e. $g_{Z(a',\Delta=1)}(z_t \mid w, \mathbf{z}_{t-1}, n_{t-1}) \equiv p(Z_t(a', \Delta = 1) = z_t \mid W = w, \mathbf{Z}_{t-1}(a', \Delta = 1) = \mathbf{z}_{t-1}, N_{t-1}(a', \Delta = 1) = n_{t-1})$. Consider an intervention which imposes the following conditional distribution $g_{a,a',\Delta=1}$ on the intervention variables $(A, Z_1, dA_{C,1}, \ldots, Z_\tau, dA_{C,\tau})$:

$$\begin{aligned}
&g_{a,a',\Delta=1}(A = a \mid W) = 1 \\
&g_{a,a',\Delta=1}(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1}) \equiv g_{Z(a',\Delta=1)}(Z_t \mid W, \mathbf{Z}_{t-1}, N_{t-1}). \\
&g_{a,a',\Delta=1}(dA_{C,t} = 0 \mid W, A, \mathbf{Z}_t, N_t, A_{C,t-1}) = 1 \qquad (21)
\end{aligned}$$

The resulting counterfactual event process is $dN_t(a, Z(g_{a,a',\Delta=1}), \Delta = 1)$. To simply notation, we will use $dN_t(a, Z(g_{a,a',\Delta=1})) \equiv dN_t(a, Z(g_{a,a',\Delta=1}), \Delta = 1)$. Denote the corresponding failure time as $T(a, Z(g_{a,a',\Delta=1}))$. This experiment can be encoded as

$$\begin{aligned}
W &= f_W(U_W) \\
A &= a \\
Z_t(g_{a,a',\Delta=1}) &= f_{Z_t}^{a',\Delta=1}(W, A = a', \mathbf{Z}_{t-1}(g_{a,a',\Delta=1}), N_{t-1}(a, Z(g_{a,a',\Delta=1})), A_{C,t-1} = 0, U_{Z_t}^{a',\Delta=1}) \\
dN_t(a, Z(g_{a,a',\Delta=1})) &= f_{dN_t}(W, A = a, \mathbf{Z}_t(g_{a,a',\Delta=1}), N_{t-1}(a, Z(g_{a,a',\Delta=1})), A_{C,t-1} = 0, U_{dN_t}) \\
dA_{C,t} &= 0 \qquad (22)
\end{aligned}$$

In other words: We first set the baseline treatment to $A = a$ and disallow right censoring throughout the study. At each visit t, given realization $\big(W, A = a, \mathbf{Z}_{t-1}(g_{a,a',\Delta=1}), N_{t-1}(a, Z(g_{a,a',\Delta=1})) = (w, a, \mathbf{z}_{t-1}, n_{t-1})\big)$, we set $Z_t(g_{a,a',\Delta=1})$ to be distributed according to $g_{Z(a',\Delta=1)}(\cdot \mid w, \mathbf{z}_{t-1}, n_{t-1})$. Recall that if death has already occurred (i.e. $n_{t-1} = 1$), then $g_{Z(a',\Delta=1)}$ will assign the degenerate value with probability 1. We then measure the response $dN_t(a, Z(g_{a,a',\Delta=1}))$ under realized history $(W = w, A = a, \mathbf{Z}_t(g_{a,a',\Delta=1}) = \mathbf{z}_t, N_{t-1}(a, Z(g_{a,a',\Delta=1})) = n_{t-1})$. Similar to section 2, this formulation presupposes external specification of the function $g_{Z(a',\Delta=1)}(z_t \mid w, \mathbf{z}_{t-1}, n_{t-1})$. If this counterfactual distribution is not known, it needs to be ascertained through a separate controlled experiment. Together with the structural equations, the variables

$$\Big(U \equiv \Big(U_W, U_A, \mathbf{U}_Z \equiv (U_{Z_t} : t), \mathbf{U}_{dN} \equiv (U_{dN_t} : t), \mathbf{U}_C \equiv (U_{dA_{C,t}} : t)\Big), U_Z^g \equiv \{(U_{Z_t}^{a',\Delta=1} : t) : a' \in \mathscr{A}\}\Big), \qquad (23)$$

define a full data random variable on an individual with distribution $P_{(U,U_Z^g)}$.

Define the *natural direct effect* as $(P(T(1, Z(g_{1,0,\Delta=1})) > t_0) - P(T(0, Z(g_{0,0,\Delta=1})) > t_0))$ and *natural indirect effect* as $(P(T(1, Z(g_{1,1,\Delta=1})) > t_0) - P(T(1, Z(g_{1,0,\Delta=1})) > t_0))$. The identification and estimation of these two effects can be approached through the study of the *mediation formula*

$$\Psi_{a,a',\Delta=1}(P_{(U,U_Z^g)}) \equiv P(T(a, Z(g_{a,a',\Delta=1})) > t_0), \qquad (24)$$

The same comments (paragraph succeeding (5)) regarding the use of these terminologies with respect to the established literature apply here.

For the remaining of this section, we focus on the identification and the efficient influence function of the mediation formula $\Psi_{a,a',\Delta=1}(P_{(U,U_Z^g)})$. The analogous results regarding natural direct and indirect effects (as well as the corresponding estimators) can be derived using same steps as in section 2 — we omit those here for conciseness.

3.1 Identification of the mediation formula

Let $dN_t(a, \mathbf{z}, \Delta = 1)$ denote the event process under a deterministic intervention which sets treatment to value $A = a$, mediators to value $\mathbf{Z} = \mathbf{z}$, and censoring indicators to $\mathbf{dA_C} = \mathbf{0}$.

Theorem 3. *Suppose the following positivity assumptions regarding the data generating distribution P_0 hold:*

P1. There exists $0 < \delta_A < 1$ such that $g_{A,0}(A \mid W) > \delta_A$, a.e. over $\mathscr{A}$;

P2. There exists $0 < \delta_Z < 1$ such that $\inf_{t \in \{1,\ldots,t_0\}} g_{Z,0}(Z_t \mid W, A = a, \mathbf{Z}_{t-1}, \tilde{T} > t-1) > \delta_Z$, a.e. over $\mathscr{Z}$;

P3. There exists $0 < \delta_N < 1$ such that $\inf_{t \in \{1,\ldots,t_0\}} 1 - Q_{dN,0}(t \mid W, A = a', \mathbf{Z}_t, \tilde{T} > t-1) > \delta_N$;

P4. There exists $0 < \delta_C < 1$ such that $\inf_{t \in \{1,\ldots,t_0\}} 1 - g_{C,0}(t \mid W, A, \mathbf{Z}_t, N_t = 0, A_{C,t-1} = 0) > \delta_C$.

Let $g_{a,a',\Delta=1}$ be an intervention distribution on $(A, \mathbf{Z}, \mathbf{dA_C})$ as defined in (21). Let $T(a, Z(g_{a,a',\Delta=1}))$ be the corresponding failure time under experiment (22). Suppose the following randomization assumptions hold for all $\mathbf{z}$*:*

I1. $(\mathbf{dN}(a', \Delta = 1), \mathbf{Z}(a', \Delta = 1)) \perp A$, given W;

I2. $(\mathbf{dN}_{j>t}(a', \Delta = 1), \mathbf{Z}_{j>t}(a', \Delta = 1)) \perp dA_{C,t}$, given W, $A = a'$, $\mathbf{Z}_t$, N_t, $A_{C,t-1} = 0$;

I3. $\mathbf{dN}(a, \mathbf{z}, \Delta = 1) \perp A$, given W;

I4. $\mathbf{dN}_{j \geq t}(a, \mathbf{z}, \Delta = 1) \perp Z_t$, given $W, A = a, \mathbf{Z}_{t-1} = \mathbf{z}_{t-1}, N_{t-1}, A_{C,t-1} = 0$;

I5. $\mathbf{dN}_{j>t}(a, \mathbf{z}, \Delta = 1) \perp dA_{C,t}$, given $W, A = a, \mathbf{Z}_t = \mathbf{z}_t, N_t, A_{C,t-1} = 0$;

I6. $\mathbf{dN}_{j\geq t}(a,\mathbf{z},\Delta=1)\perp Z_t(g_{a,a',\Delta=1})$, *given* W, $A=a$, $\mathbf{Z}_{t-1}(g_{a,a',\Delta=1})=\mathbf{z}_{t-1}$, *and* $N_{t-1}(a,Z(g_{a,a',\Delta=1}))$, $A_{C,t-1}=0$;

then, (24) can be expressed as

$$\begin{aligned}&\Psi_{a,a',\Delta=1}(P_0)\\&\equiv\sum_{w}\sum_{\mathbf{z}_{t_0}}Q_{W,0}(w)\\&\times\prod_{t=1}^{t_0}\left\{g_{Z,0}\big(z_t\mid w,A=a',\mathbf{z}_{t-1},\tilde{T}>t-1\big)\big(1-Q_{dN,0}(t\mid w,A=a,\mathbf{z}_t,\tilde{T}>t-1)\big)\right\}.\end{aligned}\tag{25}$$

In addition to the identifiability conditions of theorem 1, we now also require that right censoring be sequentially randomized.

The parameter (24) can also be identified under conditional independence conditions on the joint distribution of (U,U^g): $(\mathbf{U}_Z,\mathbf{U}_{dN})\perp U_A$ given U_W; $((\mathbf{U}_{dN})_{j>t},(\mathbf{U}_Z)_{j>t})\perp U_{dA_{C,t}}$ given $U_W,U_A,(\mathbf{U}_Z)_t$, $(\mathbf{U}_{dN})_t$, $(\mathbf{U}_C)_{t-1}$; $(\mathbf{U}_{dN})_{j\geq t}\perp U_{Z_t}$ given U_W,U_A, $(\mathbf{U}_Z)_{t-1}$, $(\mathbf{U}_{dN})_{t-1},(\mathbf{U}_C)_{t-1}$; $(\mathbf{U}_{dN})_{j\geq t}\perp U_{Z_t}^{a',\Delta=1}$, given U_W, U_A, $(\mathbf{U}_Z)_{t-1}^{a',\Delta=1}$, $(\mathbf{U}_{dN})_{t-1}$, $(\mathbf{U}_C)_{t-1}$.

Similar to section 2, the natural direct effect here can also be interpreted as a weighted average of controlled direct effect, and the total effect $P(T(1,\Delta=1)>t_0)-P(T(0,\Delta=1)>t_0)$ can again be decomposed into the sum of the natural effects.

3.2 Efficient influence function of the mediation formula

Let $\mathscr{M}$ denote a locally saturated semiparametric model containing the true data generating distribution P_0. Let $g=(g_A,g_Z,g_C)$ and $Q=(Q_W,Q_{dN})$ denote the corresponding components on a given $P\in\mathscr{M}$. The shorthand notations used in section 2 are modified to incorporate censoring variables:

$$\begin{aligned}G_Z(\mathbf{Z}_t\mid W,A)&\equiv\prod_{t'=1}^{t}g_Z\left(Z_{t'}\mid W,A,\mathbf{Z}_{t'-1},N_{t'-1}=0,A_{C,t'-1}=0\right),\\\bar{Q}_N(t\mid W,A,\mathbf{Z}_t)&\equiv\prod_{t'=1}^{t}1-Q_{dN}(t'\mid W,A,\mathbf{Z}_{t'},N_{t'-1}=0,A_{C,t'-1}=0),\\\bar{G}_C(t\mid W,A,\mathbf{Z}_t)&\equiv\prod_{t'=1}^{t}1-g_C(t'\mid W,A,\mathbf{Z}_{t'},N_{t'}=0,A_{C,t'-1}=0).\end{aligned}$$

The parameter in (25) is the following map evaluated at P_0:

$$\begin{aligned}&\Psi_{a,a',\Delta=1} : \mathscr{M} \to \mathbb{R}\\ &P \mapsto \Psi_{a,a',\Delta=1}(P) \equiv \sum_{w}\sum_{\mathbf{z}_{t_0}} Q_W(w)G_Z(\mathbf{z}_{t_0} \mid w, A=a')\bar{Q}_N(t_0 \mid w, A=a, \mathbf{z}_{t_0}).\end{aligned} \tag{26}$$

Define the *conditional mediation formula at t* as:

$$\begin{aligned}&\phi_{Z,a,a',\Delta=1}(P)(t;W,\mathbf{Z}_{t-1})\\ &\equiv \sum_{\mathbf{z}_t^{t_0}}\prod_{t'=t}^{t_0} g_Z(z_{t'} \mid W, A=a', \mathbf{Z}_{t-1}, \mathbf{z}_t^{t'-1}, \tilde{T} > t'-1)\bar{Q}_N\left(t_0 \mid W, A=a, \mathbf{Z}_{t-1}, \mathbf{z}_t^{t_0}\right),\end{aligned}$$

for $t = 1, \ldots, t_0$, and $\phi_{Z,a,a',\Delta=1}(P)(t_0+1; W, \mathbf{Z}_{t_0}) \equiv \bar{Q}_N(t_0 \mid W, A=a, \mathbf{Z}_{t_0})$. Note again that $\Psi_{a,a',\Delta=1}(P) = E_{Q_W}\left(\phi_{Z,a,a',\Delta=1}(P)(t=1;W)\right)$, and that $\phi_{Z,a,a',\Delta=1}(P)$ satisfies the recursive relation

$$\phi_{Z,a,a',\Delta=1}(P)(t-1;W,\mathbf{Z}_{t-2}) = E_{g_{Z,t-1}}\left(\phi_{Z,a,a',\Delta=1}(P)(t;W,\mathbf{Z}_{t-1}) \mid W, A=a', \mathbf{Z}_{t-2}, \tilde{T} > t-2\right).$$

Theorem 4. *Let $\Psi_{a,a',\Delta=1} : \mathscr{M} \to \mathbb{R}$ be defined as in (26). Suppose the following are true for $P \in \mathscr{M}$:*

P1. There exists $0 < \delta_A < 1$ such that $g_A(A \mid W) > \delta_A$, a.e. over $\mathscr{A}$;

P2. There exists $0 < \delta_Z < 1$ such that $\inf_{t \in \{1,\ldots,t_0\}} g_Z(Z_t \mid W, A=a, \mathbf{Z}_{t-1}, \tilde{T} > t-1) > \delta_Z$, a.e. over $\mathscr{Z}$;

P3. There exists $0 < \delta_N < 1$ such that $\inf_{t \in \{1,\ldots,t_0\}} 1 - Q_{dN}(t \mid W, A=a', \mathbf{Z}_t, \tilde{T} > t-1) > \delta_N$;

P4. There exists $0 < \delta_C < 1$ such that $\inf_{t \in \{1,\ldots,t_0\}} 1 - g_C(t \mid W, A, \mathbf{Z}_t, N_t = 0, A_{C,t-1} = 0) > \delta_C$.

The Efficient Influence Function of $\Psi_{a,a',\Delta=1}$ at P is given by

$$
\begin{aligned}
& D^*_{a,a',\Delta=1}(P)(O) \\
& = -\sum_{t=1}^{t_0} I(\tilde{T} > t-1) \left\{ \frac{I(A=a)}{g_A(a \mid W)\bar{G}_C(t-1 \mid W, A=a, \mathbf{Z}_{t-1})} \prod_{t'=1}^{t} \frac{g_Z(Z_{t'} \mid W, A=a', \mathbf{Z}_{t'-1}, \tilde{T} > t'-1)}{g_Z(Z_{t'} \mid W, A=a, \mathbf{Z}_{t'-1}, \tilde{T} > t'-1)} \right. \\
& \left. \times \sum_{\mathbf{z}_{t+1}^{t_0}} \prod_{t'=t+1}^{t_0} g_Z(z_{t'} \mid W, A=a', \mathbf{Z}_t, \mathbf{z}_{t+1}^{t'-1}, \tilde{T} > t'-1) \left(1 - Q_{dN}(t' \mid W, A=a, \mathbf{Z}_t, \mathbf{z}_{t+1}^{t'})\right) \right\} \\
& \times (dN_t - Q_{dN}(t \mid W, A=a, \mathbf{Z}_t)) \\
& + \sum_{t=1}^{t_0} I(\tilde{T} > t-1) \left\{ \frac{I(A=a')}{g_A(a' \mid W)\bar{G}_C(t-1 \mid W, A=a', \mathbf{Z}_{t-1})} \frac{\phi_{Z,a,a',\Delta=1}(P)(t+1; W, \mathbf{Z}_t) - \phi_{Z,a,a',\Delta=1}(P)(t; W, \mathbf{Z}_{t-1})}{\bar{Q}_N(t-1 \mid W, A=a', \mathbf{Z}_{t-1})} \right\} \\
& + \phi_{Z,a,a',\Delta=1}(P)(t=1; W) - E_{Q_W}(\phi_{Z,a,a',\Delta=1}(P)(t=1; W)) \qquad (27)
\end{aligned}
$$

The proof of theorem 2 can be modified to prove theorem 4 by incorporating tangent subspaces corresponding to the conditional probabilities of censoring at each time t given observed history.

The difference between the EIFs (14) and (27) is that when right censoring is present, the static treatment mechanism $g_A(A \mid W)$ is replaced by the static treatment and censoring mechanism $g_A(A \mid W)\bar{G}_C(t-1 \mid W, A, \mathbf{Z}_{t-1})$.

The robustness properties of (27) is summarized in the following lemma.

Lemma 2. *Let $\Psi_{a,a',\Delta=1}(P)$ be as defined in (26); its efficient influence function under $\mathscr{M}$ is $D^*_{a,a',\Delta=1}(P)$, as given in (27).*

Suppose for $P_0 \in \mathscr{M}$, conditions of theorem 4 hold. Then,

$$P_0 D^*_{a,a',\Delta=1}(Q, g, \Psi_{a,a',\Delta=1}(P_0)) = 0$$

if one of the following holds:

R1. $Q_{dN} = Q_{dN,0}$, *and* $\phi_{Z,a,a',\Delta=1}(P) = \phi_{Z,a,a',\Delta=1}(P_0)$.

R2. $Q_{dN} = Q_{dN,0}$, *and* $g_A = g_{A,0}$ *and* $g_C = g_{C,0}$.

R3. $(g_A, g_C, g_Z) = (g_{A,0}, g_{C,0}, g_{Z,0})$ *and* $\phi_{Z,a,a',\Delta=1}(P) = \phi_{Z,a,a',\Delta=1}(g_{Z,0}, Q_{dN})$.

The proof of this lemma is also very similar to the proof of lemma 2, since the key steps are not effected by the inverse weighting by censoring probabilities.

4. SUMMARY AND DISCUSSION

In this paper, we argued that in a survival setting with time-dependent mediators, the natural effects should be defined based on blocking only those paths from treatment to mediator that are not through the survival history of interest (We illustrate in appendix A4.1 that blocking all paths from treatment to mediator would yield parameters that are not interpretable for the goal of survival mediation). In extending the traditional formulation of the corresponding effects—where the mediator is considered an intermediate outcome— to the current setting, we encountered identifiability conditions that are too strong for the purpose of this study (appendix A4.2). As an alternative, we proposed to adopt the stochastic interventions (SI) approach of Didelez et al. (2006), where the mediator is considered an intervention variable, onto which a given distribution is enforced. The second contribution of this paper is a general semiparametric inference framework for the resulting effect parameters. More specifically, efficient influence functions under a locally saturated semiparametric model are derived, and their robustness properties are established. Note that in parameters arising from mediation analysis, the mediator densities play the role of probability weights in iterated expectations (which we referred to as conditional mediation formula, conditional natural direct and indirect effects). In many applications where the mediator densities are difficult to estimate, regression-based estimators of these iterated expectations are viable alternatives to substitution-based estimators that rely on consistent estimation of the mediator densities. We also developed the g-computation, IPTW, A-IPTW and TMLE estimators for the natural effect parameters; of these, the A-IPTW and TMLE are locally semiparametric efficient and remain unbiased under certain types of model mis-specifications.

Under the SI formulation, the treatment of interest as well as the mediator variables are regarded as intervention variables. One can obtain a total effect decomposition and the subsequent definition of natural direct and indirect effects that are analogous to those in Pearl (2001). The natural direct effect (NDE) under this formulation has an intrinsic interpretation as a weighted average of controlled direct effects (CDE), since the CDE can be considered as a deterministic intervention on the treatment and mediator variables. By regarding the mediator variables as intervention variables, the SI formulation requires external specification of a counterfactual mediator distribution. It is important to note that causal mediation, under either SI or non-SI approaches, presupposes that the mediator of interest is amenable to external manipulation. In applications where such manipulations are not conceivable, we should be cautious that causal mediation can only offer answers to purely mechanistic questions defined under hypothetical experiments.

The mediation formula and its efficient influence function presented here are applicable to general multilevel treatments. In these applications, one can still use the mediation formula to define certain direct and indirect effects of interest (e.g. through marginal structural models: Robins, Hernan, and Brumback (2000), Neugebauer and van der Laan (2007)). The efficient influence functions for those parameters can be derived using the delta method.

The setting we used in this paper is based on discrete time points. In situations where one is willing to approximate a continuous failure time by discrete time points, the methods presented here can be applied. Otherwise, formal generalizations are needed to handle the analytic subtleties in a continuous time context.

Acknowledgements

This work was supported by NIH grant 5R01AI74345-5. We would also like to thank Ivan Diaz for the literature references and helpful discussions on stochastic interventions.

APPENDIX A1

Identifying the distribution of observations under intervention $g_{a,a'}$ is an application of general identifiability results for stochastic interventions. This appendix takes a brief detour to review these results.

Stochastic Interventions over Time-Dependent Treatment

Stochastic interventions (a.k.a stochastic policies, random interventions, randomized dynamic strategies, etc.) impose pre-specified probability distributions to the intervention variables. The traditional static interventions or dynamic treatment regimes can be considered as special cases of stochastic interventions where the imposed probability distributions have mass at only one point. The identification of stochastic interventions have been addressed in the literature (e.g. Dawid and Didelez (2010), Pearl (2009), Tian (2008), Robins and Richardson (2010) , Diaz and van der Laan (2011)). For self-containment, this appendix paraphrases these results as applicable to the current setting.

Consider a general longitudinal data structure $O = (L_0, A_1, L_1, \ldots, A_K, L_K)$, for some $K > 0$, with time- ordering encoded as

$$\begin{aligned} L_0 &= f_{L_0}(U_{L_0}) \\ A_t &= f_{A_t}(L_0, \mathbf{A}_{t-1}, \mathbf{L}_{t-1}, U_{A_t}) \text{ for } t = 1, \ldots, K \\ L_t &= f_{L_t}(L_0, \mathbf{A}_t, \mathbf{L}_{t-1}, U_{L_t}) \text{ for } t = 1, \ldots, K. \end{aligned} \tag{28}$$

Let $\mathbf{A} \equiv (A_1, \ldots, A_K)$ be the intervention variables. Without any interventions, this NPSEM generates the observed data $O \sim P_0$. The likelihood of $O \sim P_0$ can be factored as

$$\begin{aligned} p_0(O) &= p_0(L_0)\prod_{t=1}^{K} p_0(A_t \mid L_0, \mathbf{A}_{t-1}, \mathbf{L}_{t-1}) p_0(L_t \mid L_0, \mathbf{A}_t, \mathbf{L}_{t-1}) \\ &\equiv Q_0(L_0)\prod_{t=1}^{K} g_0(A_t \mid L_0, \mathbf{A}_{t-1}, \mathbf{L}_{t-1}) Q_0(L_t \mid L_0, \mathbf{A}_t, \mathbf{L}_{t-1}). \end{aligned}$$

Let $g(\cdot \mid L_0, \mathbf{A}_{k-1}, \mathbf{L}_{k-1}) : \mathscr{A} \to [0,1]$ be a conditional distribution for the intervention variables A_k. We say that g is *permissible* for the NPSEM (28) if the intervention to impose distribution g on the variables $\mathbf{A}$ does not change the way the non-intervention variables respond to a given history. This intervention experiment is encoded as

$$\begin{aligned} L_0 &= f_{L_0}(U_{L_0}) \\ A_t(g) &= f^g_{A_t}(L_0, \mathbf{A}_{t-1}(g), \mathbf{L}_{t-1}(A(g)), U^g_{A_t}) \text{ for } t = 1, \ldots, K \\ L_t(A(g)) &= f_{L_t}(L_0, \mathbf{A}_t(g), \mathbf{L}_{t-1}(A(g)), U_{L_t}) \text{ for } t = 1, \ldots, K. \end{aligned} \tag{29}$$

Let $\mathscr{G}$ denote the set of permissible conditional distributions for the intervention variables $\mathbf{A}$. Let P_g denote the distribution of O_g. Note that $g_0 \in \mathscr{G}$ by definition of permissibility. The observed data is $O = O_{g_0}$ and $P_{g_0} = P_0$. For a given stochastic intervention $g^* \in \mathscr{G}$, we wish to identify P_{g^*} as a function of P_{g_0} and g^*.

For a fixed vector $\mathbf{a} = (a_1, \ldots, a_K)$, let $\mathbf{L}(\mathbf{a})$ denote the counterfactual outcomes under a static intervention that sets all $\mathbf{A} = \mathbf{a}$, i.e.

$$\begin{aligned} L_0 &= f_{L_0}(U_{L_0}) \\ A_t &= a_t \\ L_t(\mathbf{a}) &= f_{L_t}(L_0, \mathbf{A}_t = \mathbf{a}_t, \mathbf{L}_{t-1}(\mathbf{a}), U_{L_t}) \text{ for } t = 1, \ldots, K \end{aligned}$$

Theorem 5. *Given $g^* \in \mathscr{G}$, the likelihood P_{g^*} of O_{g^*} can be identified as*

$$P_{g^*}(\mathbf{l}, \mathbf{a}) = P_{g_0}(l_0)\prod_{t=1}^{K} g^*(a_t \mid l_0, \mathbf{a}_{t-1}, \mathbf{l}_{t-1}) P_{g_0}(l_t \mid l_0, \mathbf{a}_t, \mathbf{l}_{t-1}), \tag{30}$$

for all $(\mathbf{l},\mathbf{a}) \in \mathscr{L}^K \times \mathscr{A}^K$*, if the following sequential randomization assumption holds for* $g \in \{g^*, g_0\}$*:*

$$\mathbf{L}_{j\geq k}(\mathbf{a}) \perp A_k(g), \textit{ given } L_0, \mathbf{L}_{k-1}(A(g)), \mathbf{A}_{k-1}(g) = \mathbf{a}_{k-1}. \tag{31}$$

In words, if for data generated under P_{g_0} and data generated under P_{g^*}, the treatment at each stage is randomized given its past history (no unmeasured confounders), then (30) is true. This is a rephrasing of the results in the literature which stated that to identify any stochastic intervention on the distribution of $\mathbf{A}$, it suffices to identify the conditional densities of $\mathbf{L}(\mathbf{a})$ under static (a.k.a. atomic) interventions. For self-containment, we detail a proof here.

Proof. The likelihood P_{g^*} can be factorized according to the time ordering in (29) into

$$\begin{aligned} P_{g^*}(\mathbf{l},\mathbf{a}) &= P_{g^*}(l_0)\prod_{t=1}^{K} P_{g^*}(a_t \mid l_0, \mathbf{a}_{t-1}, \mathbf{l}_{t-1}) P_{g^*}(l_t \mid l_0, \mathbf{a}_t, \mathbf{l}_{t-1}) \\ &\equiv P_{g^*}(l_0)\prod_{t=1}^{K} g^*(a_t \mid l_0, \mathbf{a}_{t-1}, \mathbf{l}_{t-1}) P_{g^*}(l_t \mid l_0, \mathbf{a}_t, \mathbf{l}_{t-1}). \end{aligned}$$

Since L_0 is not affected by the intervention variables, its marginal distribution is invariant under the choice of g. Therefore, $P_{g^*}(l_0) = P_{g_0}(l_0)$. It remains to show that $P_{g^*}(l_t \mid l_0, \mathbf{a}_t, \mathbf{l}_{t-1}) = P_{g_0}(l_t \mid l_0, \mathbf{a}_t, \mathbf{l}_{t-1})$ under the SRA.

Suppose for $g \in \{g_0, g^*\}$, the assumption (31) holds. Then we can write

$$
\begin{aligned}
&P(L_t(\mathbf{a}) = l_t \mid L_0 = l_0, \mathbf{L}_{t-1}(\mathbf{a}) = \mathbf{l}_{t-1}) \\
&= \frac{P(\mathbf{L}_t(\mathbf{a}) = \mathbf{l}_t \mid L_0 = l_0)}{P(\mathbf{L}_{t-1}(\mathbf{a}) = \mathbf{l}_{t-1} \mid L_0 = l_0)} \\
&= \frac{P(\mathbf{L}_t(\mathbf{a}) = \mathbf{l}_t \mid L_0 = l_0, A_1(g) = a_1)}{P(\mathbf{L}_{t-1}(\mathbf{a}) = \mathbf{l}_{t-1} \mid L_0 = l_0, A_1(g) = a_1)} \text{by applying assumption to } k = 1 \\
&= \frac{P(\mathbf{L}_2^t(\mathbf{a}) = \mathbf{l}_2^t, L_1(A(g)) = l_1 \mid L_0 = l_0, A_1(g) = a_1)}{P(\mathbf{L}_2^{t-1}(\mathbf{a}) = \mathbf{l}_2^{t-1}, L_1(A(g)) = l_1 \mid L_0 = l_0, A_1(g) = a_1)} \\
&= \frac{P(\mathbf{L}_2^t(\mathbf{a}) = \mathbf{l}_2^t \mid L_0 = l_0, A_1(g) = a_1, L_1(A(g)) = l_1)}{P(\mathbf{L}_2^{t-1}(\mathbf{a}) = \mathbf{l}_2^{t-1} \mid L_0 = l_0, A_1(g) = a_1, L_1(A(g)) = l_1)} \\
&= \frac{P(\mathbf{L}_2^t(\mathbf{a}) = \mathbf{l}_2^t \mid L_0 = l_0, L_1(A(g)) = l_1, \mathbf{A}_2(g) = \mathbf{a}(2))}{P(\mathbf{L}_2^{t-1}(\mathbf{a}) = \mathbf{l}_2^{t-1} \mid L_0 = l_0, L_1(A(g)) = l_1, \mathbf{A}_2(g) = \mathbf{a}(2))} \text{by applying assumption to } k = 2 \\
&= \frac{P(\mathbf{L}_3^t(\mathbf{a}) = \mathbf{l}_3^t, L_2(A(g)) = l_2 \mid L_0 = l_0, L_1(A(g)) = l_1, \mathbf{A}_2(g) = \mathbf{a}(2))}{P(\mathbf{L}_3^{t-1}(\mathbf{a}) = \mathbf{l}_3^{t-1}, L_2(A(g)) = l_2 \mid L_0 = l_0, L_1(A(g)) = l_1, \mathbf{A}_2(g) = \mathbf{a}(2))} \\
&= \frac{P(\mathbf{L}_3^t(\mathbf{a}) = \mathbf{l}_3^t \mid L_0 = l_0, \mathbf{L}_2(A(g)) = \mathbf{l}_2, \mathbf{A}_2(g) = \mathbf{a}(2))}{P(\mathbf{L}_3^{t-1}(\mathbf{a}) = \mathbf{l}_3^{t-1} \mid L_0 = l_0, \mathbf{L}_2(A(g)) = \mathbf{l}_2, \mathbf{A}_2(g) = \mathbf{a}(2))} \\
&= \vdots \text{ repeat the same reasoning till } k = t - 1 \\
&= P(L_t(\mathbf{a}) = l_t \mid L_0 = l_0, \mathbf{L}_{t-1}(A(g)) = l_{t-1}, \mathbf{A}_{t-1}(g) = \mathbf{a}_{t-1}) \\
&= P(L_t(\mathbf{a}) = l_t \mid L_0 = l_0, \mathbf{L}_{t-1}(A(g)) = l_{t-1}, \mathbf{A}_t(g) = \mathbf{a}_t) \text{by applying assumption to } k = t \\
&= P(L_t(A(g)) = l_t \mid L_0 = l_0, \mathbf{L}_{t-1}(A(g)) = l_{t-1}, \mathbf{A}_t(g) = \mathbf{a}_t).
\end{aligned}
$$

Applying this result to both $g = g_0$ and $g = g^*$, it follows that

$$
\begin{aligned}
&P(L_t(A(g^*)) = l_t \mid L_0 = l_0, \mathbf{L}_{t-1}(A(g^*)) = \mathbf{l}_{t-1}, \mathbf{A}_t(g^*) = \mathbf{a}_t) \\
&= P(L_t(\mathbf{a}) = l_t \mid L_0 = l_0, \mathbf{L}_{t-1}(\mathbf{a}) = \mathbf{l}_{t-1}) \\
&= P(L_t(A(g_0)) = l_t \mid L_0 = l_0, \mathbf{L}_{t-1}(A(g_0)) = \mathbf{l}_{t-1}, \mathbf{A}_t(g_0) = \mathbf{a}_t)
\end{aligned}
$$

□

Stronger but more general identifiability conditions in terms of the joint distribution of $(\mathbf{U} = (\mathbf{U}_L \equiv (U_{L_t} : t), \mathbf{U}_A \equiv (U_{A_t} : t)), \mathbf{U}_A^g \equiv (U_{A_t}^g : t))$ are: $(\mathbf{U}_L)_{j \geq t} \perp U_{A_t}$, given $(\mathbf{U}_A)_{t-1}, (\mathbf{U}_L)_{t-1}$, and $(\mathbf{U}_L)_{j \geq t} \perp U_{A_t}^g$, given $(\mathbf{U}_A)_{t-1}^g, (\mathbf{U}_L)_{t-1}$. In particular, if g is defined so that $\mathbf{U}_A^g \perp \mathbf{U}_L$, then only randomization of A under P_0 is needed.

Corollary 3. *If g is defined such that $\mathbf{U}_A^g \perp \mathbf{U}_L$, then P_g is identified under the usual SRA*

$$\mathbf{L}_{j\geq k}(\mathbf{a}) \perp A_k, \textit{ given } L_0, \mathbf{L}_{k-1}, \mathbf{A}_{k-1} = \mathbf{a}_{k-1}.$$

APPENDIX A2

Proof of Theorem 1

Part of the proof is a simple consequence of theorem 5 in appendix A1. By definition of this intervention,

$$\begin{aligned}
&P(T(a,Z(g_{a,a'})) > t_0)\\
&= \sum_w Q_0(w) \sum_{\mathbf{z}_{t_0}} \prod_{t=1}^{t_0} g_{Z(a')}(z_t \mid w, \mathbf{z}_{t-1}, n_{t-1} = 0) P(dN_t(a, Z(g_{a,a'})) = 0 \mid W = w, \mathbf{Z}_t(g_{a,a'}) = \mathbf{z}_t, N_{t-1}(a, Z(g_{a,a'})) = 0)\\
&= \sum_w Q_0(w) \sum_{\mathbf{z}_{t_0}} \prod_{t=1}^{t_0} g_{Z(a')}(z_t \mid w, \mathbf{z}_{t-1}, n_{t-1} = 0) P(dN_t(a, \mathbf{z}) = 0 \mid W = w, N_{t-1}(a, \mathbf{z})) = 0)\\
&= \sum_w Q_0(w) \sum_{\mathbf{z}_{t_0}} \prod_{t=1}^{t_0} g_{Z(a')}(z_t \mid w, \mathbf{z}_{t-1}, n_{t-1} = 0) P(dN_t = 0 \mid W = w, A = a, \mathbf{Z}_t = \mathbf{z}_t, N_{t-1} = 0).
\end{aligned}$$

The first equality is by the definition of the stochastic intervention. The second equality is by theorem 5 and conditions I4; the third equality is implied by theorem 5 and conditions I2-I3..

Applying I1 and the usual randomization argument, $g_{Z(a')}(z_t \mid w, \mathbf{z}_{t-1}, n_{t-1} = 0) \equiv P(Z_t(a') = z_t \mid W = w, \mathbf{Z}_{t-1}(a') = \mathbf{z}_{t-1}, N_{t-1}(a') = 0)$ is identified as $g_{Z,0}(z_t \mid W = w, A = a', \mathbf{Z}_{t-1} = \mathbf{z}_{t-1}, N_{t-1} = 0)$.

Proof of Theorem 2

For any $P \in \mathscr{M}$, we may factor the likelihood according to the time ordering of (1):

$$p(O) = p_W(W) p_A(A \mid W) \prod_{t=1}^{\tau} p_{Z_t}(Z_t \mid W, A, \mathbf{Z}_{t-1}, N_{t-1}) p_{dN_t}(dN_t \mid W, A, \mathbf{Z}_{t-1}, N_t). \qquad (32)$$

For $O_j \in \{W, A, Z_t, dN_t : t = 1, \tau\}$, let $Pa(O_j)$ denote the parent of O_j (i.e. all the endogenous variables that are inputs of O_j in (1)), and let P_j denote the conditional probability of $P_{O_j}(O_j \mid Pa(O_j)$.

Let $L_0^2(P)$ denote the Hilbert space of mean zero functions of O, endowed with the covariance operator. Consider a rich class of one-dimensional parametric submodels $P(\varepsilon)$ that are generated by only fluctuating P_j. Under our model, no restrictions are imposed on the conditional probabilities P_j. As a result, given any function $S_{O_j} \in L_0^2(P)$ of $(O_j, Pa(O_j))$ with finite variance and $E_P(S_{O_j}(O_j, Pa(O_j)) \mid Pa(O_j)) = 0$, the fluctuation $P_j(\varepsilon) = (1 + \varepsilon S_{O_j}(O_j, Pa(O_j)))P_j$ is a valid one-dimensional submodel with score S_{O_j}. Therefore, the tangent subspaces corresponding to fluctuations of each P_j are given by

$$\begin{aligned}
T(P_W) &= \{S_W(W) : E_P(S_W) = 0\} \\
T(P_{A|W}) &= \{S_A(A,W) : E_P(S_A \mid W) = 0\} \\
T(P_{Z_t|Pa(Z_t)}) &= \{S_{Z_t}(Z_t, W, A, \mathbf{Z}_{t-1}, N_{t-1}) : E_P(S_{Z_t} \mid W, A, \mathbf{Z}_{t-1}, N_{t-1}) = 0\} \\
T(P_{dN_t|Pa(dN_t)}) &= \{S_{dN_t}(dN_t, W, A, \mathbf{Z}_t, N_{t-1}) : E_P(S_{dN_t} \mid W, A, \mathbf{Z}_t, N_{t-1}) = 0\}.
\end{aligned}$$

Due to the factorization in (32), $T(P_i)$ is orthogonal to $T(P_j)$ for $O_i \neq O_j$. Moreover, the tangent space $T(P)$, corresponding to fluctuations of the entire likelihood, is given by the orthogonal sum of these tangent subspaces, i.e. $T(P) = \bigoplus_j T(P_j)$, and any score $S(O) \in T(P)$ can be decomposed as $\sum_j S_{O_j}(O)$.

Under this generous definition of the tangent subspaces, any function $S(O)$ that has zero mean and finite variance under P is contained in $T(P)$. This implies in particular that any gradient for the pathwise derivative of $\Psi_{a,a'}(\cdot)$ is contained in $T(P)$, and is thus in fact the canonical gradient. Therefore, it suffices to show that $D^*_{a,a'}(\cdot)$ in (14) is a gradient for the pathwise derivative of $\Psi_{a,a'}(\cdot)$.

Consider the three summands of (14):

$$D^*_{N,a,a'}(P)(O) \equiv -\sum_{t=1}^{t_0} I(N_{t-1}=0)\left\{ \frac{I(A=a)}{g_A(a \mid W)} \prod_{t'=1}^{t} \frac{g_Z(Z_{t'} \mid W, A=a', \mathbf{Z}_{t'-1}, N_{t'-1}=0)}{g_Z(Z_{t'} \mid W, A=a, \mathbf{Z}_{t'-1}, N_{t'-1}=0)} \right.$$
$$\left. \times \sum_{\mathbf{z}_{t+1}^{t_0}} \prod_{t'=t+1}^{t_0} g_Z(z_{t'} \mid W, A=a', \mathbf{Z}_t, \mathbf{z}_{t+1}^{t'-1}, N_{t'-1}=0)\left(1 - Q_{dN}(t' \mid W, A=a, \mathbf{Z}_t, \mathbf{z}_{t+1}^{t'})\right) \right\}$$
$$\times \left(dN_t - Q_{dN}(t \mid W, A=a, \mathbf{Z}_t)\right),$$

$$D^*_{Z,a,a'}(P)(O)$$
$$\equiv \sum_{t=1}^{t_0} I(N_{t-1}=0)\left\{ \frac{I(A=a')}{g_A(a' \mid W)} \frac{\left(\phi_{Z,a,a'}(P)(t+1; W, \mathbf{Z}_t) - \phi_{Z,a,a'}(P)(t; W, \mathbf{Z}_{t-1})\right)}{\bar{Q}_N(t-1 \mid W, A=a', \mathbf{Z}_{t-1})} \right\},$$

and

$$D_{W,a,a'}(P)(O) \equiv \phi_{Z,a,a'}(P)(t=1; W) - E_{Q_W}\left(\phi_{Z,a,a'}(P)(t=1; W)\right).$$

For any $S(O) = \sum_j S_{O_j}(O) \in T(P)$, let $P_S(\varepsilon)$ denote the fluctuation of P with score S. Under appropriate regularity conditions, the pathwise derivative at P can be expressed as

$$\frac{d}{d\varepsilon}\Psi_{a,a'}(P_S(\varepsilon)) \mid_{\varepsilon=0}$$
$$= \frac{d}{d\varepsilon}\left\{ \sum_w \sum_{\mathbf{z}_{t_0}} \left(((1+\varepsilon S_W)Q_W)(w) \right.\right.$$
$$\left.\left. \times \prod_{t=1}^{t_0}((1+\varepsilon S_{Z_t})P_{Z_t})(z_t \mid w, A=a', \mathbf{z}_{t-1}, N_{t-1}=0)\,(1-((1+\varepsilon S_{dN_t})P_{dN_t})(dN_t=1 \mid w, A=a, \mathbf{z}_t, N_{t-1}=0)) \right) \right\} \mid_{\varepsilon=0}$$
$$= \sum_w \sum_{\mathbf{z}_{t_0}} \left(Q_W(w) G_Z(\mathbf{z}_{t_0} \mid w, a') \bar{Q}_N(t_0 \mid w, a, \mathbf{z}_{t_0}) \right.$$
$$\left. \times \sum_{t=1}^{t_0} \frac{-S_{dN_t}(dN_t=1, N_{t-1}=0, w, a, \mathbf{z}_t) Q_{dN}(t \mid w, a, \mathbf{z}_t, N_{t-1}=0)}{1-Q_{dN}(t \mid w, a, \mathbf{z}_t, N_{t-1}=0)} \right) \tag{33}$$
$$+ \sum_w \sum_{\mathbf{z}_{t_0}} Q_W(w) G_Z(\mathbf{z}_{t_0} \mid w, A=a') \bar{Q}_N(t_0 \mid w, a, \mathbf{z}_{t_0}) \sum_{t=1}^{t_0} S_{Z_t}(\mathbf{z}_t, w, a', N_{t-1}=0) \tag{34}$$
$$+ \sum_w \sum_{\mathbf{z}_{t_0}} S_W(w) Q_W(w) G_Z(\mathbf{z}_{t_0} \mid w, A=a') \bar{Q}_N(t_0 \mid w, a, \mathbf{z}_{t_0}). \tag{35}$$

Note firstly that for every $t = 1, \dots, t_0$,

$$
\begin{aligned}
&E_P\left(D^*_{N,a,a'}(P)(O)S_{dN_t}(N_t, W, A, \mathbf{Z}_t)\right) \\
&= \sum_w \sum_{\mathbf{z}_{t_0}} \left(Q_W(w) G_Z(\mathbf{z}_{t_0} \mid w, a') \bar{Q}_N(t_0 \mid w, a, \mathbf{z}_{t_0}) \frac{-S_{dN_t}(dN_t = 1, N_{t-1} = 0, w, a, \mathbf{z}_t) Q_{dN}(t \mid w, a, \mathbf{z}_t, N_{t-1} = 0)}{1 - Q_{dN}(t \mid w, a, \mathbf{z}_t, N_{t-1} = 0)} \right).
\end{aligned}
$$

Therefore, (33) can be written as

$$
E_P\left\{ D^*_{N,a,a}(P)(O) \left(\sum_{t=1}^{t_0} S_{dN_t}(N_t, W, A, \mathbf{Z}_t) \right) \right\}.
$$

Moreover, $D_{N,a,a'}(P)(O) \in \sum_t T(P_{dN_t|Pa(dN_t)})$ by the definition of these tangent subspaces. It thus follows from the orthogonal decomposition of $T(P)$ that

$$
\begin{aligned}
&E_P\left\{ D^*_{N,a,a}(P)(O) \left(\sum_{t=1}^{t_0} S_{dN_t}(N_t, W, A, \mathbf{Z}_t) \right) \right\} \\
&= E_P\left\{ D^*_{N,a,a}(P)(O) \left(S_W(W) + S_A(A, W) + \sum_{t=1}^{t_0} S_{Z_t}(\mathbf{Z}_t, W, A, N_{t-1}) + \sum_{t=1}^{t_0} S_{dN_t}(N_t, W, A, \mathbf{Z}_t) \right) \right\}.
\end{aligned}
$$

By similar arguments, (34) can be written as

$$
\begin{aligned}
&E_P\left\{ D^*_{Z,a,a'}(P)(O) \left(\sum_{t=1}^{t_0} S_{Z_t}(\mathbf{Z}_t, W, A, N_{t-1}) \right) \right\} \\
&= E_P\left\{ D^*_{Z,a,a'}(P)(O) \left(S_W(W) + S_A(A, W) + \sum_{t=1}^{t_0} S_{Z_t}(\mathbf{Z}_t, W, A, N_{t-1}) + \sum_{t=1}^{t_0} S_{dN_t}(N_t, W, A, \mathbf{Z}_t) \right) \right\},
\end{aligned}
$$

and (35) can be written as

$$
\begin{aligned}
&E_P\left\{ D^*_{W,a,a'}(P)(O) S_W(W) \right\} \\
&= E_P\left\{ D^*_{W,a,a'}(P)(O) \left(S_W(W) + S_A(A, W) + \sum_{t=1}^{t_0} S_{Z_t}(\mathbf{Z}_t, W, A, N_{t-1}) + \sum_{t=1}^{t_0} S_{dN_t}(N_t, W, A, \mathbf{Z}_t) \right) \right\}.
\end{aligned}
$$

Combining these results, one concludes that

$$
\begin{aligned}
&\frac{d}{d\varepsilon} \Psi_{a,a'}(P_S(\varepsilon)) \mid_{\varepsilon=0} \\
&= E_P\left\{ \left(D^*_{N,a,a'}(P)(O) + D^*_{Z,a,a'}(P)(O) + D^*_{W,a,a'}(P)(O) \right) S(O) \right\}
\end{aligned}
$$

Therefore, $D^*_{a,a'}(P) = D^*_{N,a,a'}(P) + D^*_{Z,a,a'}(P) + D^*_{W,a,a'}(P)$ is a gradient for the partial derivative of $\Psi_{a,a'}$ at P. As discussed above, under the nonparametric model, $D^*_{a,a'}(P)$ is in fact the canonical gradient.

Proof of Lemma 1

Consider the efficient influence function $D^*_{a,a'}(P, \Psi_{a,a'}(P_0))$ as a function of the functionals $(Q, g, \phi_{Z,a,a'})$ of P.

$$\begin{aligned}
& P_0 D^*_{a,a'}\left(Q, g, \phi_{Z,a,a'}, \Psi_{a,a'}(P_0)\right) \\
& = -P_0 \frac{g_{A,0}(A=a \mid W)}{g_A(a \mid W)} \sum_{\mathbf{z}_{t_0}} \Bigg(\bar{Q}_N(t_0 \mid W, A=a, \mathbf{z}_{t_0}) G_Z(\mathbf{z}_{t_0} \mid W, A=a') \\
& \times \sum_{t=1}^{t_0} \frac{G_{Z,0}(\mathbf{z}_t \mid W, a)}{G_Z(\mathbf{z}_t \mid W, a)} \frac{\bar{Q}_{N,0}(t-1 \mid W, a, \mathbf{z}_{t-1})}{\bar{Q}_N(t \mid W, a, \mathbf{z}_t)} \left(Q_{dN,0}(t \mid W, a, \mathbf{z}_t) - Q_{dN}(t \mid W, a, \mathbf{z}_t)\right) \Bigg) \qquad (36) \\
& + P_0 \frac{g_{A,0}(A=a' \mid W)}{g_A(a' \mid W)} \sum_{t=1}^{t_0} \sum_{\mathbf{z}_{t-1}} \Bigg\{ \frac{\bar{Q}_{N,0}(t-1 \mid W, A=a', \mathbf{z}_{t-1})}{\bar{Q}_N(t-1 \mid W, A=a', \mathbf{z}_{t-1})} G_{Z,0}(\mathbf{z}_{t-1} \mid W, a') \\
& \times \Bigg(\sum_{z_t} g_{Z,0}(z_t \mid W, a', \mathbf{z}_{t-1}, N_{t-1}=0) \phi_{Z,a,a'}(P)(t+1; W, \mathbf{z}_t) - \phi_{Z,a,a'}(P)(t; W, \mathbf{z}_{t-1}) \Bigg) \Bigg\} \qquad (37) \\
& + P_0 \phi_{Z,a,a'}(P)(t=1; W) - \Psi_{a,a'}(P_0) \qquad (38)
\end{aligned}$$

Suppose that $Q_{dN} = Q_{dN,0}$, and $\phi_{Z,a,a'}(P) = \phi_{Z,a,a'}(P_0)$. Then, $\phi_{Z,a,a'}(P)(t; W, \mathbf{z}_{t-1})) = E_{g_{Z,0}}(\phi_{Z,a,a'}(P)(t+1; W, \mathbf{z}_t) \mid W, A=a', \mathbf{z}_{t-1}, N_{t-1}=0))$. Therefore, each of the terms (36), (37) and (38) is exactly zero.

On the other hand, if $Q_{dN} = Q_{dN,0}$ and $g_A = g_{A,0}$, then (36) is zero, and (37) can be written as a telescopic sum. Therefore, (37) plus (38) can be expressed as

$$P_0 \sum_{t=1}^{t_0} \left(\sum_{\mathbf{z}_t} G_{Z,0}(\mathbf{z}_t \mid W, a') \phi_{Z,a,a'}(P)(t+1; W, \mathbf{z}_t) - \sum_{\mathbf{z}_{t-1}} G_{Z,0}(\mathbf{z}_{t-1} \mid W, a') \phi_{Z,a,a'}(P)(t; W, \mathbf{z}_{t-1}) \right)$$
$$+ P_0 \phi_{Z,a,a'}(P)(t=1; W) - \Psi_{a,a'}(P_0)$$
$$= P_0 \left(\sum_{\mathbf{z}_{t_0}} G_{Z,0}(\mathbf{z}_{t_0} \mid W, a') \bar{Q}_{N,0}(t_0; W, A=a, \mathbf{z}_{t_0}) - \phi_{Z,a,a'}(P)(t=1; W) \right)$$
$$+ P_0 \phi_{Z,a,a'}(P)(t=1; W) - \Psi_{a,a'}(P_0)$$
$$= 0.$$

Finally, if $g_A = g_{A,0}$, $g_Z = g_{Z,0}$, and $\phi_{Z,a,a'}(P) = \phi_{Z,a,a'}(g_{Z,0}, Q_{dN})$, then (37) is exactly zero, and (36) can be written as

$$-P_0 \sum_{\mathbf{z}_{t_0}} \left(\bar{Q}_N(t_0 \mid W, A=a, \mathbf{z}_{t_0}) G_{Z,0}(\mathbf{z}_{t_0} \mid W, A=a') \right.$$
$$\left. \times \sum_{t=1}^{t_0} \frac{\bar{Q}_{N,0}(t-1 \mid W, a, \mathbf{z}_{t-1})}{\bar{Q}_N(t \mid W, a, \mathbf{z}_t)} \left(Q_{dN,0}(t \mid W, a, \mathbf{z}_t) - Q_{dN}(t \mid W, a, \mathbf{z}_t) \right) \right). \tag{39}$$

Define $h(t \mid W, a, \mathbf{z}_t) \equiv \bar{Q}_N(t \mid W, a, \mathbf{z}_t) - \bar{Q}_N(t-1 \mid W, a, \mathbf{z}_{t-1})$. Define h_0 analogously for $\bar{Q}_{N,0}$. It follows from this definition that

$$Q_{dN}(t \mid W, a, \mathbf{z}_t) = -\frac{h(t \mid W, a, \mathbf{z}_t)}{\bar{Q}_N(t-1 \mid W, a, \mathbf{z}_{t-1})}.$$

Therefore, we may rewrite the inner sum in (39) as

$$
\begin{aligned}
&\sum_{t=1}^{t_0} \frac{\bar{Q}_{N,0}(t-1 \mid W,a,\mathbf{z}_{t-1})}{\bar{Q}_N(t \mid W,a,\mathbf{z}_t)} \left(Q_{dN,0}(t \mid W,a,\mathbf{z}_t) - Q_{dN}(t \mid W,a,\mathbf{z}_t)\right) \\
&= \sum_{t=1}^{t_0} \frac{\bar{Q}_{N,0}(t-1 \mid W,a,\mathbf{z}_{t-1})}{\bar{Q}_N(t \mid W,a,\mathbf{z}_t)} \left(\frac{h(t \mid W,a,\mathbf{z}_t)}{\bar{Q}_N(t-1 \mid W,a,\mathbf{z}_{t-1})} - \frac{h_0(t \mid W,a,\mathbf{z}_t)}{\bar{Q}_{N,0}(t-1 \mid W,a,\mathbf{z}_{t-1})} \right) \\
&= \sum_{t=1}^{t_0} \left(\frac{h(t \mid W,a,\mathbf{z}_t)\bar{Q}_{N,0}(t-1 \mid W,a,\mathbf{z}_{t-1}) - h_0(t \mid W,a,\mathbf{z}_t)\bar{Q}_N(t-1 \mid W,a,\mathbf{z}_{t-1})}{\bar{Q}_N(t \mid W,a,\mathbf{z}_t)\bar{Q}_N(t-1 \mid W,a,\mathbf{z}_{t-1})} \right) \\
&= \sum_{t=1}^{t_0} \left(\frac{\Big(\bar{Q}_N(t \mid W,a,\mathbf{z}_t) - \bar{Q}_N(t-1 \mid W,a,\mathbf{z}_{t-1})\Big)\bar{Q}_{N,0}(t-1 \mid W,a,\mathbf{z}_{t-1}) - \Big(\bar{Q}_{N,0}(t \mid W,a,\mathbf{z}_t) - \bar{Q}_{N,0}(t-1 \mid W,a,\mathbf{z}_{t-1})\Big)\bar{Q}_N(t-1 \mid W,a,\mathbf{z}_{t-1})}{\bar{Q}_N(t \mid W,a,\mathbf{z}_t)\bar{Q}_N(t-1 \mid W,a,\mathbf{z}_{t-1})} \right) \\
&= \sum_{t=1}^{t_0} \left(\frac{\bar{Q}_N(t \mid W,a,\mathbf{z}_t) - \bar{Q}_{N,0}(t \mid W,a,\mathbf{z}_t)}{\bar{Q}_N(t \mid W,a,\mathbf{z}_t)} - \frac{\bar{Q}_N(t-1 \mid W,a,\mathbf{z}_{t-1}) - \bar{Q}_{N,0}(t-1 \mid W,a,\mathbf{z}_{t-1})}{\bar{Q}_N(t-1 \mid W,a,\mathbf{z}_{t-1})} \right) \\
&= \frac{\bar{Q}_N(t_0 \mid W,a,\mathbf{z}_{t_0}) - \bar{Q}_{N,0}(t_0 \mid W,a,\mathbf{z}_{t_0})}{\bar{Q}_N(t_0 \mid W,a,\mathbf{z}_{t_0})} - \frac{\bar{Q}_N(0 \mid W,a) - \bar{Q}_{N,0}(0 \mid W,a)}{\bar{Q}_N(0 \mid W,a)} \\
&= \frac{\bar{Q}_N(t_0 \mid W,a,\mathbf{z}_{t_0}) - \bar{Q}_{N,0}(t_0 \mid W,a,\mathbf{z}_{t_0})}{\bar{Q}_N(t_0 \mid W,a,\mathbf{z}_{t_0})}.
\end{aligned}
$$

Hence, (39) is simplified into

$$
\begin{aligned}
&-P_0 \sum_{\mathbf{z}_{t_0}} \left(\bar{Q}_N(t_0 \mid W, A=a, \mathbf{z}_{t_0}) G_{Z,0}(\mathbf{z}_{t_0} \mid W, A=a') \frac{\bar{Q}_N(t_0 \mid W,a,\mathbf{z}_{t_0}) - \bar{Q}_{N,0}(t_0 \mid W,a,\mathbf{z}_{t_0})}{\bar{Q}_N(t_0 \mid W,a,\mathbf{z}_{t_0})} \right) \\
&= \Psi_{a,a'}(P_0) - P_0 \sum_{\mathbf{z}_{t_0}} G_{Z,0}(\mathbf{z}_{t_0} \mid W, A=a') \bar{Q}_N(t_0 \mid W,a,\mathbf{z}_{t_0}) \\
&= \Psi_{a,a'}(P_0) - P_0 \phi_{Z,a,a'}(g_{Z,0}, Q_{dN})(t=1;W),
\end{aligned}
$$

which cancels with (38).

APPENDIX A3

In this appendix, we consider estimators for the NIE parameter (9). Analogous to the NDE case in section 2.3, an estimator $\hat{\phi}_{Z,NIE,n}(\cdot)$ of $\phi_{Z,NIE}(g_{Z,0},\cdot)$ maps an estimator $\hat{Q}_{dN,n}$ of $Q_{dN,0}$ to an estimator $\hat{\phi}_{Z,NIE,n}(\hat{Q}_{dN,n})$ of $\phi_{Z,NIE}(g_{Z,0},Q_{dN,0})$. This estimating procedure can be plug-in or regression-based. For a plug-in estimator, $\hat{\phi}_{Z,NIE,n}(\hat{Q}_{dN,n}) \equiv \phi_{Z,NIE}(\hat{g}_{Z,n},\hat{Q}_{dN,n})$. For a regression-based estimator, $\hat{\phi}_{Z,NIE,n}(\hat{Q}_{dN,n})(t;W,A,\mathbf{Z}_{t-1})$ regresses $\bar{\hat{Q}}_{N,n}(t_0 \mid W, A=1, \mathbf{Z}_{t_0})$ on $(W,A,\mathbf{Z}_{t-1})$ among observations that haven't failed by time $t-1$.

The tools for defining the g-computation and A-IPTW estimators are readily provided in the main section. We will only discuss the IPTW and TMLE estimators here.

IPTW

Consider the following function:

$$D_{NIE,IPTW}(P) = \frac{I(A=1)}{g(1 \mid W)} \left(1 - \prod_{t=1}^{t_0} \frac{g_Z(Z_t \mid W, A=0, \mathbf{Z}_{t-1}, N_{t-1})}{g_Z(Z_t \mid W, 1, \mathbf{Z}_{t-1}, N_{t-1})} \right) I(N_{t_0}=0).$$

Note firstly that $I(N_{t_0}=0)$ if and only if $I(N_t=0)$ for $t=1,\ldots,t_0)$. Next,

$$\begin{aligned}
&P_0 D_{NIE,IPTW}(P_0) = \\
&= E_{Q_{W,0}} \left(\sum_{\mathbf{z}_{t_0}} \prod_{t=1}^{t_0} g_{Z,0}(z_t \mid W, A=1, \mathbf{z}_{t-1}, N_{t-1}=0) \left(1 - Q_{dN,0}(t \mid W, A=1, \mathbf{z}_t, N_{t-1}=0)\right) \right. \\
&\left. - \sum_{\mathbf{z}_{t_0}} \prod_{t=1}^{t_0} g_{Z,0}(z_t \mid W, A=0, \mathbf{z}_{t-1}, N_{t-1}=0) \left(1 - Q_{dN,0}(t \mid W, A=1, \mathbf{z}_t, N_{t-1}=0)\right) \right) \\
&= \Psi_{NIE}(P_0)
\end{aligned}$$

Given estimators $\hat{g}_{A,n}$ and $\hat{g}_{Z,n}$ of $g_{A,0}$ and $g_{Z,0}$, respectively, an estimator of the natural indirect effect can be obtained using $D_{NIE,IPTW}$:

$$\hat{\Psi}_{NIE}^{IPTW}(P_n) \equiv \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i=1)}{\hat{g}_{A,n}(1 \mid W_i)} \left(1 - \prod_{t=1}^{t_0} \frac{\hat{g}_{Z,n}(Z_{i,t} \mid W_i, A=0, \mathbf{Z}_{i,t-1}, N_{i,t-1})}{\hat{g}_{Z,n}(Z_{i,t} \mid W_i, A=1, \mathbf{Z}_{i,t-1}, N_{i,t-1})} \right) I(N_{i,t_0}=0)$$

As noted earlier, if $N_{i,t-1} \neq 0$, we assign $\hat{g}_{Z,n}(Z_{i,t} \mid W_i, A, \mathbf{Z}_{i,t-1}, N_{i,t-1})$ the value of 1. This way the estimator is well-defined.

TMLE

To construct the TMLE estimator for the natural indirect effect (13), we first decompose the EIF (16) into its components on the tangent subspaces:

$$
\begin{aligned}
&D^*_{NIE,N}(P) \equiv \sum_{t=1}^{t_0} D^*_{NIE,dN_t}(P) \\
&= -\sum_{t=1}^{t_0} I(N_{t-1}=0) \frac{I(A=1)}{g_A(1 \mid W) G_Z(\mathbf{Z}_t \mid W, A=1))} \\
&\times \left\{ \sum_{\mathbf{z}_{t+1}^{t_0}} \left(\left(G_Z\left(\mathbf{Z}_t, \mathbf{z}_{t+1}^{t_0} \mid W, A=1\right) - G_Z\left(\mathbf{Z}_t, \mathbf{z}_{t+1}^{t_0} \mid W, A=0\right)\right) \prod_{t'=t+1}^{t_0} 1 - Q_{dN}(t' \mid W, A=1, \mathbf{Z}_t, \mathbf{z}_{t+1}^{t'}) \right) \right\} \\
&\times (dN_t - Q_{dN}(t \mid W, A=1, \mathbf{Z}_t)), \\
&D^*_{NIE,Z}(P) \equiv \sum_{t=0}^{t_0} D^*_{NIE,Z_t}(P) \\
&= \sum_{t=1}^{t_0} I(N_{t-1}=0) \left\{ \frac{2A-1}{g_A(A \mid W)} \frac{1}{\bar{Q}_N(t-1 \mid W, A, \mathbf{Z}_{t-1})} (\phi_{Z,NIE}(P)(t+1; W, A, \mathbf{Z}_t) - \phi_{Z,NIE}(P)(t; W, A, \mathbf{Z}_{t-1})) \right\}, \\
&D^*_{NIE,W}(P) \\
&= (\phi_{Z,NIE}(P)(t=1; W, A=1) - \phi_{Z,NIE}(P)(t=1; W, A=0)) - E_{Q_W}(\phi_{Z,NIE}(P)(t=1; W, A=1) - \phi_{Z,NIE}(P)(t=1; W, A=0)).
\end{aligned}
$$

We note that the empirical marginal distribution $\hat{Q}_{W,n}$ of W is a consistent estimator of $Q_{W,0}$ that readily satisfies the equation $P_n D^*_{NIE,W}(\phi_{Z,NIE}(P), \hat{Q}_{W,n}) = 0$ for any $\phi_{Z,NIE}(P)$. Hence, the proposed estimator will focus on targeted estimation of $Q_{dN,0}$, and $\phi_{Z,NIE}(P_0)$.

To simplify notation, we will use $Q_{dN}(t)$ to denote $Q_{dN}(t \mid W, A=1, \mathbf{Z}_t, N_{t-1})$. We use the minus loglikelihood function

$$L_{dN_t}(Q_{dN}(t))(O) = I(N_{t-1}=0) I(A=1) \log\left(Q_{dN}(t)^{dN_t}(1-Q_{dN}(t))^{1-dN_t}\right).$$

Under this loss function, consider the logistic working submodel

$$Q_{dN}(t)(\varepsilon) \equiv expit\,(logit\,(Q_{dN}(t)) + \varepsilon C_{dN}(g, Q_{dN})(t)),$$

where

$$
\begin{aligned}
&C_{dN}(g,Q_{dN})(t)\\
&\equiv \frac{1}{g_A(1 \mid W)G_Z(\mathbf{Z}_t \mid W,A=1)}\\
&\times \sum_{\mathbf{z}_{t+1}^{t_0}} \left(\left(G_Z\left(\mathbf{Z}_t,\mathbf{z}_{t+1}^{t_0} \mid W,A=1\right) - G_Z\left(\mathbf{Z}_t,\mathbf{z}_{t+1}^{t_0} \mid W,A=0\right)\right) \prod_{t'=t+1}^{t_0} 1 - Q_{dN}(t' \mid W,A=1,\mathbf{Z}_t,\mathbf{z}_{t+1}^{t'}) \right)
\end{aligned}
\tag{40}
$$

It is suppressed in this notation that $C_{dN}(g,Q_{dN})(t)$ is a function of $(W,\mathbf{Z}_t)$. Note that for a given t, $C_{dN}(g,Q_{dN})(t)$ only depends on $Q_{dN}(j)$ for $j>t$.

Recall the recursive relation

$$\phi_{Z,NIE}(P)(t;W,A,\mathbf{Z}_{t-1}) = E_{g_{Z,t}}\left(\phi_{Z,NIE}(P)(t+1;W,A,\mathbf{Z}_t) \mid W,A,\mathbf{Z}_{t-1},N_{t-1}=0\right).$$

We suppress the notation $\phi_{Z,NIE}(P)(t;W,A,\mathbf{Z}_{t-1})$ into $\phi_{Z,NIE}(P)_t$. Consider the following loss function for $\phi_{Z,NIE}(P)_t$:

$$
\begin{aligned}
&L_{Z_t}\left(\phi_{Z,NIE}(P)_t\right)\\
&\equiv -I(N_{t-1}=0)\log\left(\left(\phi_{Z,NIE}(P)(t)\right)^{\phi_{Z,NIE}(P)(t+1)} \left(1-\phi_{Z,NIE}(P)(t)\right)^{1-\phi_{Z,NIE}(P)(t+1)} \right),
\end{aligned}
$$

with parametric working submodel given by

$$\phi_{Z,NIE}(P)_t(\varepsilon) = expit\left(logit\left(\phi_{Z,NIE}(P)_t\right) + \varepsilon C_Z(g,Q_{dN})(t)\right),$$

where

$$C_Z(g,Q_{dN})_t = \frac{2A-1}{g_A(A|W)\bar{Q}_N(t-1 \mid W,A,\mathbf{Z}_{t-1})}. \tag{41}$$

Implementation

Let $\hat{g}_{A,n}$, $\hat{Q}_{dN,n}$ and $\hat{g}_{Z,n}$, be initial estimators of $g_{A,0}$, $Q_{dN,0}$ and $g_{Z,0}$, respectively.

1. Starting with $t=t_0$, $C_{dN}(\hat{g}_n,\hat{Q}_{dN,n})(t_0) \equiv \frac{1}{\hat{g}_{A,n}(1|W)\hat{G}_{Z,n}(\mathbf{Z}_{t_0}|W,A=1)}$. Obtain an ε for $\hat{Q}_{dN,n}(t_0)$ given by

$$\hat{\varepsilon}^*_{dN,t_0} = \arg\min_{\varepsilon} P_n L_{dN_{t_0}}\left(\hat{Q}_{dN,n}(t_0)(\varepsilon)\right).$$

This provides an TMLE estimate $\hat{Q}^*_{dN,n}(t_0) \equiv \hat{Q}_{dN,n}(t_0)(\hat{\varepsilon}^*_{dN,t_0})$ for $Q_{dN}(t_0 \mid W, A = 1, \mathbf{Z}(t_0))$.

2. For each $1 \leq t < t_0$, let $\hat{Q}^*_{dN,n}$ denote the vector of TMLE estimates

$$\hat{Q}^*_{dN,n} \equiv (\hat{Q}^*_{dN,n}(t_0), \hat{Q}^*_{dN,n}(t_0 - 1), \ldots, \hat{Q}^*_{dN,n}(t+1))$$

obtained thus far. We use these and $\hat{G}_{Z,n}$ to construct $C_{dN}(\hat{g}_n, \hat{Q}^*_{dN,n})(t)$ as prescribed in (40). The optimal ε for $\hat{Q}_{dN,n}(t)$ is thus given by $\hat{\varepsilon}^*_{dN,t} = \arg\min_\varepsilon P_n L_{dN_t}\left(\hat{Q}_{dN,n}(t)(\varepsilon)\right)$.

The step 2 above updates $\hat{Q}_{dN,n}(t)$ sequentially in the order of descending t. Once we have obtained all the t_0 updates, let $\hat{Q}^*_{dN,n}$ to represent the TMLE estimator for the function $Q_{dN,0}$ at times $t = 1, \ldots, t_0$.

Let $\hat{\phi}_{Z,NIE,n}(\cdot)$ be an estimating procedure for $\phi_{Z,NIE}(P_0)$ which maps the TMLE estimator $\hat{Q}^*_{dN,n}$ to an estimator $\hat{\phi}_{Z,NIE,n}(\hat{Q}^*_{dN,n})$ of $\phi_{Z,NIE}(P_0)$. Our next steps will update $\hat{\phi}_{Z,NIE,n}(\cdot)$ towards optimal bias-variance tradeoff for the parameter of interest. We suppress the notation $\hat{\phi}_{Z,NIE,n}(\hat{Q}^*_{dN,n})(t; W, A, \mathbf{Z}_{t-1})$ into $\hat{\phi}_{Z,NIE,n}(t)$.

3. Define

$$\hat{\phi}^*_{Z,NIE,n}(t_0 + 1; W, A, \mathbf{Z}_{t_0}) \equiv \hat{\bar{Q}}^*_{N,n}(t_0 \mid W, A = 1, \mathbf{Z}_{t_0})$$

4. For each $1 \leq t \leq t_0$, suppose we have obtained the TMLE estimator $\hat{\phi}^*_{Z,NIE,n}(t+1)$ for $\phi_{Z,NIE}(P_0)(t+1; W, A, \mathbf{Z}_t)$. The parametric submodel $\hat{\phi}_{Z,NIE}(t)(\varepsilon)$ is constructed using $C_Z(\hat{g}, \hat{Q}^*_{dN,n})(t)$ as given in (41), and the optimal ε is given by

$$\hat{\varepsilon}^*_{Z_t} = \arg\min_\varepsilon P_n L_{Z_t}\left(\hat{\phi}_{Z,NIE,n}(t)(\varepsilon)\right).$$

This yields the TMLE estimator $\hat{\phi}^*_{Z,NIE,n}(t) \equiv \hat{\phi}_{Z,NIE,n}(t)(\hat{\varepsilon}^*_{Z_t})$ of $\phi_{Z,NIE}(P_0)(t; W, A, \mathbf{Z}_{t-1})$.

5. We perform the updates in step 4 sequentially in order of decreasing t. Once, we have obtained the TMLE estimator $\hat{\phi}^*_{Z,NIE,n}(t = 1; W, A)$, the TMLE estimate

of the natural indirect effect is given by

$$\hat{\Psi}^{TMLE}_{NIE}(P_n) \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{\phi}^{*}_{Z,NIE,n}(t=1;W_i,A=1) - \hat{\phi}^{*}_{Z,NIE,n}(t=1;W_i,A=0).$$

This estimator is substitution-based and the relevant components of the likelihood ($Q_{dN,0}$ and $\phi_{Z,NIE}(P_0)$) are estimated so that $P_n D^{*}_{NIE}(\hat{g}_n, \hat{Q}^{*}_{dN,n}, \hat{\phi}^{*}_{Z,NIE,n}, \hat{\Psi}^{TMLE}_{NIE}) = 0$. Therefore, this estimator also inherit the robustness properties of corollary 2.

APPENDIX A4

In this appendix, we evaluate the various options to formulate the causal mediation problem in the survival setting with time-dependent mediator, without regarding mediators as intervention variables. The first option is a simple extension of the traditional natural effects definition in the existing literature (e.g. van der Laan and Petersen (2004), VanderWeele (2010), Robins and Richardson (2010), Tchetgen Tchetgen and VanderWeele (2012)), where all the paths from the treatment to the mediators are blocked. We shall see that the resulting ideal experiment is not well-defined for the purpose of mediating the effect on the event process. The second option leaves the paths from treatment to mediator through survival history unblocked. However, the sufficient identifiability conditions, while reasonable in other applications, may be too strong for survival study. As a result, we argue that a SI-based perspective of causal mediation offers an attractive alternative to formulate the effect parameters.

We begin by reviewing the one time-point setting. Under the non-SI approach introduced by Robins and Greenland (1992) and Pearl (2001), one defines a counterfactual event indicator $dN(a,Z(a'))$ according to the following experiment

$$\begin{aligned} W &= f_W(U_W) \\ Z(a') &= f_Z(W, A=a', U_Z), \\ dN(a,Z(a')) &= f_{dN}(W, A=a, Z(a'), U_{dN}). \end{aligned} \tag{42}$$

$dN(a,Z(a'))$ is the event indicator in an ideal experiment where A is set to a, and the intermediate variable Z takes its value under the influence of $A = a'$. The identifiability conditions (Pearl (2001)) for $P(dN(a,Z(a')) = 0)$ are $dN(a,z) \perp (A,Z) \mid W$, $dN(a,z) \perp Z(a') \mid W$, and $Z(a') \perp A \mid W$.

A4.1: Blocking all paths from treatment to mediators

A direct extension of (42) is to conceptualize the mediator process as being defined entirely in a world with $A = a'$. The counterfactual survival time of interest would be $T(a,\mathbf{Z}(a')) = T(A = a, \mathbf{dN}(A = a), \mathbf{Z}(A = a', \mathbf{dN}(A = a')))$. The hypothetical experiment generating this survival time is:

$$\begin{aligned} W &= f_W(U_W) \\ A &= a \\ Z_t(a') &= f_{Z_t}(W, A = a', \mathbf{Z}_{t-1}(a'), N_{t-1}(a'), U_{Z_t}), \\ dN_t(a') &= f_{dN_t}(W, A = a', \mathbf{Z}_t(a'), N_{t-1}(a'), U_{dN_t}), \\ dN_t(a,\mathbf{Z}(a')) &= f_{dN_t}(W, A = a, \mathbf{Z}_t(a'), N_{t-1}(a,\mathbf{Z}(a')), U_{dN_t}). \end{aligned} \tag{43}$$

The experiment can be run by either drawing variables subsequently according to the order above, or by first drawing $(\mathbf{Z}(a'), \mathbf{dN}(a'))$, and then draw $\mathbf{dN}(a,\mathbf{z})$ with the given realization of $\mathbf{Z}(a') = \mathbf{z}$. Either way, if one draws $dN_t(a') = 1$, i.e. event happens at time t under treatment $A = a'$, the next mediator $Z_{t+1}(a')$ is assigned the degenerate value at time $t+1$. Then, drawing $dN_{t+1}(a,\mathbf{Z}(a'))$, when the latest mediator has the degenerate value but $N_t(a, \mathbf{Z}(a')) = 0$, is not defined. One could deterministically set $dN_{t+1}(a, \mathbf{Z}(a')) = 1$ in this case and still obtain a well-defined survival time, but this would allow the effect of treatment $A = a'$ on survival to influence the effect of treatment $A = a$ on survival, which is contrary to the purpose of mediating the effect of $A = a$ on the survival process.

In this light, well-defined mediation formulas and natural effects in the current setting should not block the paths of treatment to mediator through the survival history. In other words, the direct effect questions should be rephrased to ”what is the effect of treatment on survival, if treatment had no other effect on the mediators other than through the survival history?”.

A4.2: Only blocking those paths from treatment to mediators that are not through survival history

Due to the considerations above, we wish to define mediation effects where the paths from treatment to mediator through the outcome process is left unblocked. These effects of interest are extension of the path-specific effects discussed in Pearl (2001), Avin et al. (2005) and Robins and Richardson (2010). Consider the following hypothetical experiment:

$$\begin{aligned}
W &= f_W(U_W) \\
A &= a \\
Z_t(a', N(a)) &\equiv f_{Z_t}(W, A = a', \mathbf{Z}_{t-1}(a', N(a)), N_{t-1}(a, Z(a')), U_{Z_t}), \\
dN_t(a, Z(a')) &\equiv f_{dN_t}(W, A = a, \mathbf{Z}_{t-1}(a', N(a)), N_{t-1}(a, Z(a')), U_{dN_t}). \qquad (44)
\end{aligned}$$

Note the simplified notation for the counterfactuals $Z_t(a', N(a))$ and $dN_t(a, Z(a'))$: for instance, at $t = 2$, $Z_2(a', N(a))$ is in fact $Z_2(a', N_1(a, Z_1(a')))$, and $dN_2(a, Z(a'))$ is in fact $dN_2(a, (Z_2(a', N_1(a, Z_1(a'))), Z_1(a')))$. This experiment differs from the (43) in that the event process affecting each mediator response is the outcome process of interest. More specifically, under (44) the experiment first sets $A = a$; at each visit, given realization $(W = w, A = a, \mathbf{Z}_{t-1}(a', N(a)) = \mathbf{z}_{t-1}, N_{t-1}(a, Z(a')) = n_{t-1})$, it measures the response Z_t would have had if the treatment were $A = a'$ while the rest of the history remained the same; then, with given realization $(W = w, A = a, \mathbf{Z}_t(a', a) = \mathbf{z}_t, N_{t-1}(a, Z(a')) = n_{t-1})$, it measures the event indicator dN_t. Abusing the notation, let $T(a, a')$ denote the survival time resulting from (44).

The difference between the experiment in (44) and the SI-based experiment in (4) lies in that under the SI formulation, the conditional probability $P(Z_t(g_{a,a'}) = z_t \mid W, A = a, \mathbf{Z}_{t-1}(g_{a,a'}), N_{t-1}(a, Z(g_{a,a'})))$ is known (intervened) to be the unspecified function $g_{Z(a')}(z_t \mid W, \mathbf{Z}_{t-1}(g_{a,a'}), N_{t-1}(a, Z(g_{a,a'})))$, therefore, one only needs to identify the function $g_{Z(a')}$. Under (44), the conditional probability $P(Z_t(a', N(a)) \mid W, \mathbf{Z}_{t-1}(a', N(a)), N_{t-1}(a, Z(a')))$ is not known and remains to be identified. Therefore, even though the SI-based parameter $P(T(a, Z(g_{a,a'})) > t_0)$ and this non-SI parameter $P(T(a, a') > t_0)$ would identify to the same statistical parameter (6), they are differently formulated causal parameters.

$P(T(a, a') > t_0)$ is identified to (6) if the following independences hold for the distribution of U: $(\mathbf{U}_Z, \mathbf{U}_{dN}) \perp U_A$ given U_W, $(\mathbf{U}_{dN})_{j \geq t} \perp U_{Z_t}$ given $U_W, U_A, (\mathbf{U}_Z)_{t-1}, (\mathbf{U}_{dN})_{t-1}$, and $(\mathbf{U}_Z)_{j>t} \perp U_{dN_t}$ given $U_W, U_A, (\mathbf{U}_Z)_t, (\mathbf{U}_{dN})_{t-1}$. The last assumption would imply that the event indicator at a given time t be independent of future potential mediators — this condition is too strong for the purpose of effect mediation in a survival study. We note, however, that these are only sufficient conditions for identification, whether weaker conditions are possible for (44) (perhaps under additional assumptions on the model), is an important topic of investigation.

REFERENCES

Avin, C., I. Shpitser, and J. Pearl (2005): “Identifiability of path-specific effects,” .

Bembom, O. and M. van der Laan (2008): “Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators,” *UC Berkeley Division of Biostatistics Working Paper Series*, 230.

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1997): *Efficient and Adaptive Estimation for Semiparametric Models*, Springer-Verlag.

Dawid, A. and V. Didelez (2010): “Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview,” *Statistics Surveys*, 4, 184–231.

Diaz, I. and M. van der Laan (2011): “Population intervention causal effects based on stochastic interventions.” *Biometrics*.

Didelez, V., A. Dawid, and S. Geneletti (2006): “Direct and indirect effects of sequential treatments,” in *Proceedings of the 22nd Annual Conference on Uncertainty in Artifical Intelligence*, 138–146.

Hernan, M., B. Brumback, and J. Robins (2000): “Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men,” *Epidemiology*, 11, 561–570.

Holland, P. (1986): “Statistics and causal inference,” *Journal of the American Statistical Association*, 81, 945–960.

Imai, K., L. Keele, and T. Yamamoto (2010): “Identification, inference and sensitivity analysis for causal mediation effects,” *Statistical Science*, 25, 51–71.

Lange, T. and J. Hansen (2011): “Direct and indirect effects in a survival context,” *Epidemiology*, 22, 575.

Neugebauer, R. and M. van der Laan (2007): "Nonparametric causal effects based on marginal structural models," *Journal of Statistical Planning and Inference*, 137, 419–434.

Pearl, J. (2001): "Direct and indirect effects," in *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, Citeseer, 411–420.

Pearl, J. (2009): *Causality: Models, Reasoning and Inference*, New York: Cambridge University Press, 2nd edition.

Pearl, J. (2011): "The mediation formula: A guide to the assessment of causal pathways in nonlinear models," in C. Berzuini, P. Dawid, and L. Bernardinelli, eds., *Causality: Statistical Perspectives and Applications*.

Petersen, M., S. Sinisi, and M. van der Laan (2006): "Estimation of direct causal effects." *Epidemiology*, 17, 276–284.

Robins, J. (1997): "Causal inference from complex longitudinal data," in E. M. Berkane, ed., *Latent Variable Modeling and Applications to Causality*, Springer Verlag, New York, 69–117.

Robins, J. (1999): "Marginal structural models versus structural nested models as tools for causal inference," in *Statistical models in epidemiology: the environment and clinical trials*, Springer-Verlag, 95–134.

Robins, J. (2003): "Semantics of causal dag models and the identification of direct and indirect effects," in N. H. P. Green and S. Richardson, eds., *Highly Structured Stochastic Systems*, Oxford University Press, Oxford, 70–81.

Robins, J. and S. Greenland (1992): “Identifiability and exchangeability for direct and indirect effects,” *Epidemiology*, 3, 143–155.

Robins, J. and A. Rotnitzky (2001): “Comment on the Bickel and Kwon article, “Inference for semiparametric models: Some questions and an answer”,” *Statistica Sinica*, 11, 920–936.

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

Robins, J. M. and T. S. Richardson (2010): “Alternative graphical causal models and the identification of direct effects.” in P. Shrout, ed., *In Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, Oxford University Press.

Rosenbaum, P. and D. B. Rubin (1983): “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, 70, 41–55.

Rubin, D. (1978): “Bayesian inference for causal effects: the role of randomization,” *Annals of Statistics*, 6, 34–58.

Stitelman, O., V. D. Gruttola, C. Wester, and M. van der Laan (2011): “Rcts with time-to-event outcomes and effect modification parameters,” in M. J. van der Laan and S. Rose, eds., *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer.

Tchetgen Tchetgen, E. (2011): “On causal mediation analysis with a survival outcome,” *The International Journal of Biostatistics*, 7, 1–38.

Tchetgen Tchetgen, E. J. and T. J. VanderWeele (2012): "On identification of natural direct effects when a confounder of the mediator is directly affected by exposure," http://biostats.bepress.com/harvardbiostat/paper148: Harvard University Biostatistics Working Paper Series.

Tein, J. and D. MacKinnon (2003): "Estimating mediated effects with survival data," *New Developments on Psychometrics. Tokyo, Japan: Springer-Verlag Tokyo Inc*, 405–12.

Tian, J. (2008): "Identifying dynamic sequential plans," in *Proceedings of the Twenty-Fourth Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)(D. McAllester and A. Nicholson, eds.)*, 554–561.

van der Laan, M. and M. Petersen (2004): "Estimation of direct and indirect causal effects in longitudinal studies," Technical report 155, Division of Biostatistics, University of California, Berkeley.

van der Laan, M. and J. Robins (2003): *Unified methods for censored longitudinal data and causality*, Springer, New York.

van der Laan, M. and S. Rose (2011): *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, Springer, first edition.

van der Laan, M. and D. Rubin (2006): "Targeted maximum likelihood learning," *The International Journal of Biostatistics*, 2.

VanderWeele, T. (2010): "Direct and indirect effects for neighborhood-based clustered and longitudinal data," *Sociological Methods & Research*, 38, 515–544.

VanderWeele, T. (2011): "Causal mediation analysis with survival data," *Epidemiology*, 22, 582.