Submitted to the Statistical Science

Structural nested models and G-estimation: the partially realized promise

Stijn Vansteelandt and Marshall Joffe

Abstract. Structural nested models (SNMs) and the associated method of G-estimation were first proposed by James Robins over two decades ago as approaches to modeling and estimating the joint effects of a sequence of treatments or exposures. The models and estimation methods have since been extended to dealing with a broader series of problems, and have considerable advantages over the other methods developed for estimating such joint effects. Despite these advantages, the application of these methods in applied research has been relatively infrequent; we view this as unfortunate. To remedy this, we provide an overview of the models and estimation methods as developed, primarily by Robins, over the years. We provide insight into their advantages over other methods, and consider some possible reasons for failure of the methods to be more broadly adopted, as well as possible remedies. Finally, we consider several extensions of the standard models and estimation methods.

Key words and phrases: Causal effect, Confounding, Direct effect, Instrumental variable, Mediation, Time-varying confounding.

1. INTRODUCTION

Structural nested models (SNMs) were designed in part to deal with confounding by variables affected by treatment (Robins, 1986). The problem arises when one is interested in estimating the joint effect of a sequence of treatments in the presence of a variable L with three characteristics, depicted in Figure 1:

1. It is independently associated with the outcome Y of interest. This can happen because (1a) it is a direct cause of the outcome, or because (1b) it shares unmeasured common causes with the outcome of interest.
2. It predicts subsequent levels (A_1) of the treatment;
3. It is affected by earlier treatment (A_0).

As a motivating example, consider an observational study of the effect of erythropoietin alpha (EPO) on mortality in a population with end-stage renal disease (ESRD) receiving hemodialysis. Patients on dialysis tend to be anemic, as

Stijn Vansteelandt is Professor of Statistics, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, B-9000 Gent, Belgium. (e-mail: stijn.vansteelandt@UGent.be) Marshall Joffe is Professor of Biostatistics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, U.S.A. (e-mail: mjoffe@mail.med.upenn.edu)

commonly measured via hematocrit (Hct) or hemoglobin levels. EPO is used to treat the anemia and stimulate the body's production of red blood cells; Hct (L) thus satisfies covariate characteristic 3. Furthermore, patients with more severe anemia (lower Hct) typically receive higher doses of EPO (characteristic 2), and sicker patients tend to be more anemic (characteristic 1(b)). Both these characteristics 1 and 2 make Hct a confounder of the effect of later treatment, requiring adjustment to estimate the effect of EPO A_1. Observational studies of the effect of extended EPO dosing on mortality will thus be characterized by confounding by a variable (Hct) affected by treatment.

In settings like the above, where the interest lies in estimating the joint effect of a sequence of treatments, standard methods which attempt to estimate these effects simultaneously (e.g., regression of Y on A_0 and A_1 or some function of both) will be inappropriate, whether or not one adjusts for or conditions on the confounder L. Characteristics 1a and 3 make Hct (L) an intermediate variable on the pathway from early EPO treatment A_0 to outcome Y; adjustment for it blocks the path $A_0 \rightarrow L \rightarrow Y$, making it impossible to find the part of the effect of early EPO treatment (A_0) mediated by Hct. Characteristics 1b and 3 make Hct (L) a so-called collider (Pearl, 1995) on the path $A_0 \rightarrow L \leftarrow U \rightarrow Y$; conditioning on or adjusting for it induces associations between A_0 and Y even if no effect of A_0 on Y exists.

Figure 1 'Causal diagram for time-varying treatment' about here

Over an extended period of time, James Robins (with some help from collaborators) introduced three basic approaches for dealing with such confounding: the parametric G-formula (Robins, 1986), structural nested models (Robins, 1989; Robins et al., 1992) with the associated method of G-estimation, and marginal structural models (Robins, Hernan and Brumback, 2000) with the associated method of inverse probability of treatment weighting. As we will argue throughout this paper, SNMs and G-estimation are, in principle, better tailored for dealing with failure of the usual assumptions of no unmeasured confounders or sequential ignorability often used to justify the application of all of these methods, as well as with (near) positivity violations whereby certain strata contain (nearly) no treated or untreated subjects (Robins, 2000). Despite these advantages, the application of these methods in applied research has been relatively infrequent.

Broadly speaking, there are two types of SNMs: models for the effect of a treatment or sequence of treatments on the mean of an outcome, and models for the effect of a treatment on the entire distribution of the outcome(s). The former include structural nested mean models (SNMMs), which have close links to structural nested cumulative failure time models (SNCFTMs) for survival outcomes; the latter include structural nested distribution models (SNDMs), which have close links to structural nested failure time models (SNFTMs) for survival outcomes. For pedagogic purposes, we will introduce these models first for point treatments (i.e., treatments which are administered at one specific time point) in Section 2. We then discuss identifying assumptions and the associated G-estimation method in Section 3, and contrast it with alternative estimation methods for the effect of a point treatment in Section 4. These results are extended to time-varying treatments in Sections 5 and 6. We show how to predict the effects of interventions in Section 7, examine extensions to mediation analysis in Section

8 and conclude with a discussion.

2. STRUCTURAL MODELS FOR POINT TREATMENTS

2.1 Structural Mean Models

Let Y^a denote the outcome in a given subject that would be seen were the subject to receive treatment a. This variable is a potential outcome, which we connect to the observed outcome through the consistency assumption that $Y = Y^a$ if the observed treatment $A = a$; otherwise, Y^a is counterfactual. Causal effects can now be defined as comparisons of potential outcomes Y^a and $Y^{a^\dagger}$ for the same individual subject or group of subjects for different treatments a and $a^\dagger$ (Rubin, 1978; Robins, 1986). In particular, letting $a^\dagger = 0$ for notational convenience, average causal effects can be defined in terms of comparisons of average potential outcomes, e.g. $E(Y^a \mid L = l, A = a) - E(Y^0 \mid L = l, A = a)$ or $E(Y^a \mid L = l, A = a)/E(Y^0 \mid L = l, A = a)$.

Structural Mean Models (SMMs) (Robins, 1994, 2000) parameterize average causal effects in subjects receiving level a of treatment as

$$g\left\{E(Y^a \mid L = l, A = a)\right\} - g\left\{E(Y^0 \mid L = l, A = a)\right\} = \gamma^*(l, a; \psi^*), \tag{1}$$

for all l and a. Here, $g(.)$ is a known link function (e.g. the identity, log or logit link), $\gamma^*(l, a; \psi)$ is a known function, smooth in ψ and satisfying $\gamma^*(l, 0; \psi) = 0$ for all l and ψ. Here and throughout, ψ^* is the true unknown finite-dimensional parameter. With $a = 0$ encoding absence of treatment - as we will assume throughout - SMMs thus express the effect of removal of treatment on the outcome mean.

Typically the parameterisation is chosen to be such that $\gamma^*(l, a; 0) = 0$ for all a and l, so that $\psi = 0$ encodes the null hypothesis of no treatment effect. For instance, for scalar covariate L one may consider the additive or linear SMM (which uses the identity link $g(x) = x$):

$$E(Y^a \mid L = l, A = a) - E(Y^0 \mid L = l, A = a) = (\psi_0^* + \psi_1^* l)a, \tag{2}$$

for unknown ψ_0^*, ψ_1^*. With A a binary exposure coded as 1 for treatment and 0 for no treatment, ψ_0^* thus encodes the average treatment effect in the treated with covariate value $L = 0$, and ψ_1^* measures how much the average treatment effect in the treated differs between subgroups with a unit difference in L. Likewise, the multiplicative or loglinear SMM uses the log link $g(x) = \log(x)$, e.g.:

$$\frac{E(Y^a \mid L = l, A = a)}{E(Y^0 \mid L = l, A = a)} = \exp\left\{(\psi_0^* + \psi_1^* l)a\right\},$$

and the logistic SMM uses the logit link $g(x) = \text{logit}(x)$, e.g.:

$$\frac{\text{odds}(Y^a = 1 \mid L = l, A = a)}{\text{odds}(Y^0 = 1 \mid L = l, A = a)} = \exp\left\{(\psi_0^* + \psi_1^* l)a\right\},$$

where $\text{odds}(V = 1 \mid W) \equiv P(V = 1 \mid W)/P(V = 0 \mid W)$ for random variables V and W. If treatment A can take on more than two values, then - without additional assumptions - the function $\gamma^*(l, a; \psi^*)$ cannot be interpreted simply as a dose response function. This is because a dose response would contrast outcomes in the same subset at different levels of a (i.e., contrast $E(Y^a \mid L = l, A = a)$

with $E(Y^{a'} \mid L = l, A = a)$ for $a \neq a'$), whereas the functions $\gamma^*(l, a; \psi^*)$ and $\gamma^*(l, a'; \psi^*)$ for $a \neq a'$ contrast causal effects for two different groups (namely, those with $A = a$ versus $A = a'$, but the same $L = l$). We will revisit this subtlety in Section 7.

One can use a SMM to construct a variable $U^*(\psi)$ whose mean value (in a subset of individuals with given covariates and treatment) equals the mean outcome that would have been seen had treatment been removed from that subset. Let

$$U^*(\psi) \equiv Y - \gamma^*(L, A; \psi),$$

if $g(.)$ is the identity link,

$$U^*(\psi) \equiv Y \exp\left\{-\gamma^*(L, A; \psi)\right\},$$

if $g(.)$ is the log link and

$$U^*(\psi) \equiv \text{expit}\left[\text{logit}\left\{E\left(Y \mid L, A\right)\right\} - \gamma^*(L, A; \psi)\right], \tag{3}$$

if $g(.)$ is the logit link. Then

$$E\left\{U^*(\psi^*) \mid L, A\right\} = E\left(Y^0 \mid L, A\right). \tag{4}$$

This identity will be central to the estimation methods for ψ^* that we will describe in Section 3. We could have defined $U^*(\psi)$ in general - and in particular for the identity and log link - as $U^*(\psi) \equiv g^{-1}\left[g\{E(Y|L, A)\} - \gamma^*(L, A; \psi)\right]$. This makes the definition of $U^*(\psi)$ dependent on the expectation $E(Y|L, A)$, which can be undesirable when this demands additional modelling. This (or some alternative) is necessary for the logit link. Special estimation methods will therefore be required for logistic SMMs.

SMMs can also be used to describe the effect of a multivariate point treatment. For instance, for a bivariate treatment $A = (A^{(1)}, A^{(2)})'$, one may use a SMM with $\gamma^*(L, A; \psi) = \psi_1 A^{(1)} + \psi_2 A^{(2)} + \psi_3 A^{(1)} A^{(2)}$ to describe the effect of setting both treatments to zero. When primary interest lies in the interaction (ψ_3) between $A^{(1)}$ and $A^{(2)}$ in their effect on the outcome, then one may instead consider the class of less restrictive Structural Mean Interaction Models (Vansteelandt et al., 2008a; Tchetgen Tchetgen, 2012). To guard against misspecification of the main treatment effects, these further relax the SMM restrictions by merely parameterising the contrast between the effects of $A^{(1)}$ when $A^{(2)}$ is set to some value $a^{(2)}$ versus zero (or of the effects of $A^{(2)}$ when $A^{(1)}$ is set to some value $a^{(1)}$ versus 0):

$$\begin{aligned} & g\left\{E\left(Y^{a^{(1)}, a^{(2)}} \mid A = a, L = l\right)\right\} - g\left\{E\left(Y^{0, a^{(2)}} \mid A = a, L = l\right)\right\} \\ & - g\left\{E\left(Y^{a^{(1)}, 0} \mid A = a, L = l\right)\right\} + g\left\{E\left(Y^{0,0} \mid A = a, L = l\right)\right\} \\ & = \gamma^*(l, a^{(1)}, a^{(2)}; \psi^*), \end{aligned} \tag{5}$$

for $a = (a^{(1)}, a^{(2)})'$; here, $\gamma^*(l, a^{(1)}, a^{(2)}; \psi)$ is a known function which encodes the interaction between both treatments, and which must be smooth in ψ and satisfy $\gamma^*(l, 0, a^{(2)}; \psi) = \gamma^*(l, a^{(1)}, 0; \psi) = 0$ for all $l, a^{(1)}, a^{(2)}$ and ψ. For instance, the natural choice $\gamma^*(l, a^{(1)}, a^{(2)}; \psi) = \psi a^{(1)} a^{(2)}$ imposes that the interaction between both exposures is the same at all levels of l.

2.2 Structural Distribution Models

When the outcome mean does not adequately summarize the data or the interest lies more broadly in evaluating treatment effects on the outcome distribution, then Structural Distribution Models (SDMs) can be used instead. These are closely related to SMMs, but instead map percentiles y of the conditional distribution of Y^a, given $L = l$ and $A = a$, into percentiles $\gamma(y, l, a; \psi^*)$ of the conditional distribution of Y^0, given $L = l$ and $A = a$. In particular, they postulate that

$$F_{Y^0|L=l,A=a}\{\gamma(y, l, a; \psi^*)\} = F_{Y^a|L=l,A=a}(y), \tag{6}$$

for all l and a. As with SMMs, $\gamma(y, l, a; \psi)$ is a known function, smooth in ψ and satisfying $\gamma(y, l, 0; \psi) = y$ for all y, l. With $a = 0$ encoding the absence of treatment, SDMs thus express the effect of removing treatment on the outcome distribution rather than the outcome mean.

Typically the parameterisation of a SDM is chosen to be such that $\gamma(y, l, a; 0) = y$, so that $\psi = 0$ encodes the null hypothesis of no treatment effect. For instance, for scalar covariate L, one could assume that

$$F_{Y^0|L=l,A=a}(y - \psi_0^* a - \psi_1^* al) = F_{Y^a|L=l,A=a}(y), \tag{7}$$

for all l and a. This characterises a location shift model following which the conditional distribution of Y^0, given L and A, can be obtained by shifting the conditional distribution of Y, given L and A, by $-\psi_0^* A - \psi_1^* AL$. One can use this to construct a variable

$$U(\psi) \equiv \gamma(Y, L, A; \psi^*)$$

whose distribution (in a subset of individuals with given covariates and treatment) is the same as that of the outcome that would have been seen had treatment been removed from that subset, in the sense that

$$F_{Y^0|L,A}(y) = F_{U(\psi^*)|L,A}(y); \tag{8}$$

e.g., $U(\psi) = Y - \psi_0 A - \psi_1 AL$ in the location shift example. This will be useful for the estimation of ψ^*.

SDMs have a stronger variant called rank preserving SDMs (Robins and Tsiatis, 1991), which postulate that

$$Y^0 = \gamma(Y, L, A; \psi^*).$$

For instance, a stronger variant of the location shift model of the previous paragraph assumes that $Y^0 = Y - \psi_0^* A - \psi_1^* AL$. By making a mapping between the potential outcomes themselves (rather than between distributions), such rank preserving SDMs are easier to understand and communicate. However, they are seldom plausible because they impose that the rankings of two subjects with different outcome values but identical treatment and covariates are preserved after mapping into Y^0 (hence the term 'rank preserving'). In particular, they assume that subjects with identical outcome, treatment and covariate values experience identical treatment effects.

Location shift SDMs like (7) make substantially stronger assumptions than correspondingly parameterized SMMs. The distribution models assume that treatment level a shifts each percentile of the conditional distribution of Y, given $L =$

$l, A = a$ by a value $\gamma^*(l, a; \psi^*)$ constant for all y (i.e., $\gamma(y, l, a; \psi) = y - \gamma^*(l, a; \psi)$), whereas the mean model assumes only a mean shift of $\gamma^*(l, a; \psi)$. When location shift is implausible, it can sometimes be made more plausible by transforming y. For instance, for strictly positive y, one might obtain a location shift SDM by defining $\gamma(y, l, a; \psi) = \exp\{\log(y) - \gamma(l, a; \psi)\}$. There will then be a correspondence between the parameters of the SDM and those of a SMM for $\log(y) - \log\{\gamma(y, l, a; \psi)\}$.

The parameterization and interpretation of SDMs that are not simply shift models can be tricky. This is because, by the nature of the cumulative distribution function, the function $\gamma(y, l, a; \psi)$ must be increasing in y for each l, a and ψ, and it may be difficult to impose that. For instance, the function $\gamma(y, a, l; \psi) = y - a\psi_1 - ya\psi_2$ may appear natural, but is not guaranteed increasing in y. An alternative function which is naturally increasing in y is $\gamma(y, a, l; \psi) = y\exp(-a\psi_2) - a\psi_1$. Here, interpretation is somewhat subtle; while ψ_2 expresses the effect of treatment A on the residual variability of Y, it also has implications for the effect of treatment on the mean of Y, and so ψ_1 cannot be interpreted simply as the effect of treatment on the mean outcome, unless $\psi_2 = 0$.

SDMs lend themselves naturally to the analysis of failure times. For instance, consider model (6) with T^a, T^0 and t substituting for Y^a, Y^0 and y. Then the choice $\gamma(t, a, l; \psi) = t\exp(a\psi_0 + al\psi_1)$ implies the failure time model defined by:

$$S_{T^0|L=l,A=a}\{t\exp(-a\psi_0^* - al\psi_1^*)\} = S_{T|L=l,A=a}(t),$$

for all l and a, where $S(\cdot)$ denotes the survival function. This model, which is an example of a Structural Accelerated Failure Time Model (SAFTM) (Robins, 1989; Robins and Tsiatis, 1991; Robins, 1992; Robins et al., 1992), expresses that treatment lengthens lifetime by a factor $\exp(a\psi_0^* + al\psi_1^*)$ (in distribution) among subjects with treatment a and covariate l.

2.3 Structural mean and distribution models for repeated measures outcomes

Structural mean and distribution models require some modification for repeated measures outcomes. The modifications for SMMs are simpler, but also allow a new class of models for discrete-time failures. Extension of SDMs is more complicated. We consider these in order.

We begin with some notation common to both types of models. Suppose that measurements on exposure and confounders are collected at time point t_0 and that outcome measurements are recorded at fixed later time points $t_1, ..., t_{K+1}$. Let for a variable X, X_k denote the level of the variable that one obtains at time t_k. We use overbars to denote the history of a variable; thus, $\overline{X}_k = \{X_0, X_1, ..., X_k\}$ denotes the history of X through t_k. We use underbars to denote the future of a variable; thus, $\underline{X}_k \equiv \{X_k, \ldots, X_{K+1}\}$. Finally, we use $\underline{X}$ as shorthand notation for $\underline{X}_1$ and $X_{k:m}$ for $m \geq k$ to denote $(X_k, ..., X_m)$.

2.3.1 Structural Mean Models and Structural Cumulative Failure-time Models Extension of SMMs to repeated measures is relatively straightforward, because they model separately the effect of a treatment on each component outcome. SMMs parameterize contrasts of $\underline{Y}^a$ and $\underline{Y}^0$ as

$$g\{E(\underline{Y}^a \mid L = l, A = a)\} - g\left\{E(\underline{Y}^0 \mid L = l, A = a)\right\} = \gamma^*(l, a; \psi^*),$$

for all l and a. Here, $g(.)$ is a known $(K+1)$-dimensional link function, $\gamma^*(l,a;\psi)$ is a known $(K+1)$-dimensional function with components $\gamma_k^*(l,a;\psi), k=1,...,K+1$, that parameterize the effect of treatment on Y_k. These components are smooth in ψ and satisfy $\gamma_k^*(l,0;\psi)=0$ for all l and ψ. For instance, the SMM defined by

$$E(Y_k^a \mid L=l, A=a) - E(Y_k^0 \mid L=l, A=a) = (\psi_0^* + \psi_1^* l)a(t_k - t_0),$$

for $k=1,...,K+1$, expresses that the effect of treatment a may depend on covariates l and changes linearly over time, being zero at the baseline time t_0.

Under this repeated measures SMM, as in Section 2.1, it is possible to define a transformation $U^*(\psi)$ of the observed outcome vector $\underline{Y}$ so that

$$E\left\{U^*(\psi^*) \mid L, A\right\} = E(\underline{Y}^0 \mid L, A).$$

Here, $U^*(\psi)$ is a vector with components $Y_k - \gamma_k^*(L,A;\psi)$ for $k=1,...,K+1$ if $g(.)$ is the identity link, $Y_k \exp\left\{-\gamma_k^*(L,A;\psi)\right\}$ if $g(.)$ is the log link, and $\text{expit}\left[\text{logit}\left\{E\left(Y_k \mid L, A\right)\right\} - \gamma_k^*(L,A;\psi)\right]$ if $g(.)$ is the logit link.

Structural Cumulative Failure Time Models (SCFTMs; Picciotto et al., 2012) are a variant of repeated measures loglinear SMMs for the modelling of cumulative failure time probabilities:

$$\frac{P\left(T^a < t_k \mid L=l, A=a\right)}{P\left(T^0 < t_k \mid L=l, A=a\right)} = \exp\left\{\gamma_k^*(l,a;\psi^*)\right\},$$

for all l, a and $k=1,...,K+1$. A limitation of this class of models is that their parameterization can be tricky when the cumulative probability of failure becomes large, because the model does not restrict the outcome probabilities to stay below 1. Martinussen et al. (2011) independently proposed a continuous-time version of this model and lay out connections with additive hazard models.

2.3.2 Structural Distribution Models For multivariate outcomes, SDMs parameterize the effect of a treatment A on the marginal distribution of the vector of future potential outcomes $\underline{Y}^a$. This mapping is typically done recursively, taking the components Y_k^a and Y_k in forward sequence. These models are therefore most easily understood by first considering the class of more restrictive rank-preserving SDMs, which postulate that, for subjects with $A=a$ and $L=l$:

$$Y_k^0 = \gamma_k(Y_k, \overline{Y}_{k-1}, l, a; \psi^*) \tag{9}$$

for $k=1,...,K+1$. Here, $\gamma_k(y_k, \overline{y}_{k-1}, l, a; \psi)$ is a known function, smooth in ψ and monotonic in y_k, and $\gamma_k(y_k, \overline{y}_{k-1}, l, 0; \psi) = y_k$ for all $\overline{y}_k, l$, and ψ. For instance, with two time points $(K=1)$, a rank preserving SDM may be given by the following set of restrictions:

$$\begin{aligned} Y_2^0 &= Y_2 - (\psi_1^* + \psi_2^* Y_1)A \\ Y_1^0 &= Y_1 - \psi_3^* A. \end{aligned} \tag{10}$$

If the effect of A on Y_2 varies with Y_1, as in this example, then one must model this explicitly since the model would otherwise - perhaps unrealistically - assume that treatment does not affect the correlation between repeated outcomes (conditional on A, L). This is unlike in SMMs where one can average the effect of A on Y_2 over

all Y_1-values. This makes it substantially more difficult to parameterize SDMs than SMMs. It moreover complicates the interpretation of effects; e.g. ψ_1^* in (10) is difficult to interpret when $\psi_2^* \neq 0$ since it expresses the effect of treatment on Y_2 in subjects with $A = 1$ and $Y_1 = 0$, where Y_1 may itself be affected by treatment. Equation (10) may hence by easier to interpret upon re-expressing it as

$$Y_2^0 = Y_2 - \left\{\psi_1^* + \psi_2^*(Y_1^0 + \psi_3^* A)\right\} A = Y_2 - \left(\psi_1^* + \psi_2^* Y_1^0 + \psi_2^* \psi_3^* A\right) A.$$

A SDM relaxes the restrictions of the rank-preserving SDM by demanding that the equality (9) merely holds in distribution, conditional on $L = l$ and $A = a$. Assuming that given L and A, $\underline{Y}$ has a continuous multivariate distribution with probability 1, a SDM can thus be defined by the set of restrictions

$$F_{\underline{Y}^0 \,|L=l,A=a}\left\{\gamma(\underline{y}, l, a; \psi^*)\right\} = F_{\underline{Y}^a \,|L=l,A=a}\left(\underline{y}\right) = F_{\underline{Y} \,|L=l,A=a}\left(\underline{y}\right),$$

for all l, a, where

$$\gamma(\underline{y}, l, a; \psi^*) \equiv \{\gamma_1(y_1, l, a; \psi^*), \gamma_2(\overline{y}_2, l, a; \psi^*), ..., \gamma_{K+1}(\overline{y}_{K+1}, l, a; \psi^*)\}$$

is given by (Robins, Rotnitzky and Scharfstein, 2000):

$$\begin{aligned} \gamma_1(y_1, l, a; \psi^*) &= F^{-1}_{Y_1^0 \,|L=l,A=a} \circ F_{Y_1 \,|L=l,A=a}(y_1) \\ \gamma_k(\overline{y}_k, l, a; \psi^*) &= F^{-1}_{Y_k^0 \,|L=l,A=a,Y_{k-1}^0=\gamma_{1:k-1}(\overline{y}_{k-1},l,a;\psi^*)} \\ &\quad \circ F_{Y_k \,|L=l,A=a,\overline{Y}_{k-1}=\overline{y}_{k-1}}(y_k), \end{aligned}$$

for $k = 2, ..., K+1$. For instance, the SDM corresponding to (10) may be written:

$$\begin{aligned} F_{Y_1^0 \,|L=l,A=a}\,(y_1 - \psi_3^* a) &= F_{Y_1 \,|L=l,A=a}\,(y_1) \\ (11)\quad F_{Y_2^0 \,|L=l,A=a,Y_1^0=y_1-\psi_3^* a}\,\{y_2 - (\psi_1^* + \psi_2^* y_1)a\} &= F_{Y_2 \,|L=l,A=a,Y_1=y_1}\,(y_2)\,. \end{aligned}$$

The decomposition of the causal effects in the blip functions $\gamma_k(\overline{y}_k, l, a; \psi^*)$ is recursive because one must model not merely average effects but instead the full mapping between distributions. In particular, the effect of treatment on the first potential outcome is modeled first; then, mapping between distributions is done successively for the outcome at successive times. The overall blip function encoded by $\gamma(.)$ and the first element of this function has the usual structure and interpretation of causal estimands; i.e., as a comparison of distributions of potential outcomes under different interventions for the same group of subjects. However, the component functions $\gamma_k(.), k > 1$ do not in general have this interpretation, since the conditioning in these mapping functions is not common between Y_k^0 and Y_k; for instance, the lefthand side of (11) conditions on Y_1^0, whereas the righthand side conditions on Y_1. Nonetheless, these component functions are causal in the sense that they represent the impact of treatment on the conditional distribution of a variable. This feature is shared with the causal rate or hazard ratio (Hernan, 2010). Under the strong assumption of rank preservation, the conditioning is on a common variable, and so then the components of the blip function do have a standard causal interpretation.

For repeated measures outcomes, SDMs correspond with similarly parameterized SMMs if the SDMs are shift models. In a shift SDM, the component functions $\gamma_k(y_k, \overline{y}_{k-1}, l, a)$ may be written as $\gamma_k(y_k, \overline{y}_{k-1}, l, a) = y_k - \gamma_k^*(l, a; \psi)$. These require that the shift in percentiles of the distribution of y_k not only be independent of y_k but also of $\overline{y}_{k-1}$. Thus, shift SDMs make substantially stronger assumptions than similarly parameterized SMMs.

Under the SDM, a $K+1$-dimensional variable $U(\psi^*) = \{U_1(\psi^*), ..., U_{K+1}(\psi^*)\}$ can be constructed with components $U_k(\psi) = \gamma_k(\overline{Y}_k, L, A; \psi)$. This vector mimics the counterfactual outcome vector $\underline{Y}^0$ in the sense that

$$P\left\{U(\psi^*) > \underline{y}|L, A\right\} = P\left(\underline{Y}^0 > \underline{y}|L, A\right\}.$$

This result will be useful for estimation.

2.4 Retrospective blip models

The blip functions and causal models discussed above largely consider the effect of a blip of treatment conditional only on treatment and baseline covariates; the sole exception has been SDMs for repeated measures outcomes, where the effect of treatment on later outcomes is modeled additionally conditional on earlier outcomes, and where the interpretation of the model parameters is not clear as a usual causal contrast. This focus is consistent with an orientation of the models to be more directly useful for making decisions, where the effect of treatment is modeled conditional only on information available at the time of the decision.

For explanatory purposes, one can construct structural models for the effect of a treatment conditional on information not available at the time of treatment. Such models may have explanatory uses even though the quantities they model are less directly relevant for making decisions. Consider modeling the effect of screening mammography on breast cancer mortality (Joffe, Small and Hsu, 2007). To a first approximation, one might assume that the mammogram has an effect on death only among subjects for whom it detects a tumor. Suppose that some subjects undergo screening at the start of the study ($A = 1$; $A = 0$ otherwise). Let L_1 indicate 1 if cancer is detected at time t_1 after the start of the study and 0 otherwise. It is of interest to know how much the screening mammogram affects mortality for subjects for whom it is effective in detecting cancer. We can then model the effect of the treatment on the outcome using a retrospective SDM (RSDM) or SFTM, which conditions on L_1 in addition to treatment and baseline covariates:

$$F_{Y^0|L_0=l_0,L_1=l_1,A=a}\left\{\gamma(y, l_0, l_1, a; \psi^*)\right\} = F_{Y|L_0=l_0,L_1=l_1,A=a}(y). \tag{12}$$

In this example, we might assume that $\gamma(y, l_0, 0, a; \psi^*) = y$ to reflect that screening has no effect in subjects for whom no tumor is detected. Note that though L_1 may be affected by A, conditioning on it does not distort the interpretation of the parameters as encoding a causal effect because identity (12) still involves a comparison of the same subjects (those with $L_0 = l_0, L_1 = l_1, A = a$) under different interventions.

Models of this sort might also be useful in determining whether the effect of a treatment given at baseline is modified by post-treatment covariates and so whether there are identifiable subgroups of subjects for whom treatment appears

not to be working (Stephens, Keele and Joffe, 2013). Changes or additions to treatment might then be proposed in such subgroups after baseline. Joffe, Small and Hsu (2007) consider the relation between these retrospective models and the popular approach of principal stratification (Frangakis and Rubin, 2002). These models can generalize to a sequence of time-varying treatments, where there are additional justifications for their use (see Section 5.3).

3. IDENTIFICATION AND ESTIMATION IN STRUCTURAL MODELS FOR POINT TREATMENTS

Two kinds of assumptions have been proposed for use in most of the literature on estimation in SMMs and SDMs: no unmeasured confounders and instrumental variables type assumptions. In this section, we will focus on the former, and defer discussion of the latter to Section 6.3.

3.1 Ignorability

The required no unmeasured confounders assumption for the identification of the parameter ψ^* indexing SMMs and SDMs can be formulated as

$$A \perp\!\!\!\perp Y^0 \mid L, \tag{13}$$

where $U \perp\!\!\!\perp V \mid W$ for random variables U, V, W denotes that U is conditionally independent of V, given W. This assumption, which is empirically unverifiable, expresses that L is sufficient to adjust for confounding of the association between A and Y. Assumption (13), which is also referred to as the weak ignorability or exchangeability assumption, is weaker than the strong ignorability assumption of Rosenbaum and Rubin (1984) which, for binary treatments, states that $A \perp\!\!\!\perp (Y^0, Y^1)|L$. However, it is generally difficult to imagine settings where assumption (13) holds, but strong ignorability fails (one exception might be settings where individuals choose treatment on the basis of their perceived belief of benefit, which may be correlated with actual benefit $Y^1 - Y^0$). That (13) is a weaker assumption is exhibited in the fact that, for binary treatments, it only identifies the effect of treatment on the treated - a contrast that has been of interest in econometrics and epidemiology (Greenland and Robins, 1986):

$$\begin{aligned} E\left(Y^1 - Y^0 \mid A=1, L\right) &= E\left(Y^1 \mid A=1, L\right) - E\left(Y^0 \mid A=1, L\right) \\ &= E\left(Y^1 \mid A=1, L\right) - E\left(Y^0 \mid A=0, L\right) \\ &= E\left(Y \mid A=1, L\right) - E\left(Y \mid A=0, L\right); \end{aligned}$$

the second equality follows due to ignorability (13) and the third due to the consistency assumption. The parameters of SMMs, SDMs and SCFTMs represent the effect of treatment in the treated (or, more generally, the effect of receiving treatment level a for subjects who received level a of treatment), and so this weaker assumption is sufficient for identification.

It follows by a similar reasoning that the blip functions in the SMMs and SDMs discussed in Sections 2.1-2.3 are nonparametrically just identified under ignorability (Robins, Rotnitzky and Scharfstein, 2000). That is, the contrast of the outcomes under the observed treatment and the outcomes that would have been seen in the absence of treatment is computable for each level of a and l (and,

for SDMs, of y) from the law of the observables without assuming any restrictions or parameterization on these functions. While such nonparametric identification is of limited use in complex settings (especially with time-varying treatments considered subsequently), due to the curse of dimensionality (Robins and Ritov, 1997), it does ensure the ability to check the assumptions in any assumed causal model (provided a sufficient sample size). In contrast, the retrospective blip functions considered in Section 2.4 are not identified nonparametrically (Vansteelandt, 2010; Stephens, Keele and Joffe, 2013). Multiple retrospective blip models may thus explain the same law of the observables equally well even under ignorability.

3.2 Estimation under ignorability

The SMM together with the ignorability assumption (13) implies that

$$E\left\{U^*(\psi^*) \mid L, A\right\} = E\left(Y^0 \mid A, L\right) = E\left(Y^0 \mid L\right) = E\left\{U^*(\psi^*) \mid L\right\}.$$

Estimation of ψ^* in a SMM can thus be based on solving estimating equations:

$$0 = \sum_{i=1}^{n} \left[d^*(A_i, L_i) - E\left\{d^*(A_i, L_i) \mid L_i\right\}\right]\left[U_i^*(\psi) - E\left\{U_i^*(\psi) \mid L_i\right\}\right], \tag{14}$$

which essentially set the empirical conditional covariance between $U^*(\psi)$ and arbitrary functions $d^*(A, L)$ of the dimension of ψ, given L, to zero. For instance, for model (2), the choice $d^*(A_i, L_i) = (1\ L_i)' A_i$ results in estimating equations

$$0 = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ L_i \end{pmatrix} \left\{A_i - E(A_i \mid L_i)\right\}\left[Y_i - E(Y_i \mid L_i) - (\psi_0 + \psi_1 L_i)\left\{A_i - E(A_i \mid L_i)\right\}\right], \tag{15}$$

from which estimates for (ψ_0, ψ_1) can be solved. A locally efficient estimator of ψ^* (under the SMM together with the ignorability assumption (13)) can be attained by setting

$$d^*(A, L) = E\left\{\frac{\partial U^*(\psi^*)}{\partial \psi} \mid A, L\right\},$$

when the variance of $U^*(\psi^*)$ given A, L is constant; local here means that the efficiency is only attained when this constant variance assumption is met and models for all conditional expectations involved in (14) are correctly specified.

The SDM together with the ignorability assumption (13) implies the more restrictive constraint that

$$U(\psi^*) \perp\!\!\!\perp A \mid L. \tag{16}$$

This motivates estimating ψ^* by picking the value ψ that makes this conditional independence hold. This forms the default approach in SAFTMs, where estimation is based on a grid search whereby the independence (16) is tested for different values of ψ^* using a (standard) statistical test until it is found to be satisfied (Robins et al., 1992). Equivalently, estimation can be based on solving an estimating equation of the form

$$\begin{aligned} 0 = &\sum_{i=1}^{n} d\left\{U_i(\psi), A_i, L_i\right\} - E\left[d\left\{U_i(\psi), A_i, L_i\right\} \mid L_i, U_i(\psi)\right] \\ &- E\left(d\left\{U_i(\psi), A_i, L_i\right\} - E\left[d\left\{U_i(\psi), A_i, L_i\right\} \mid L_i, U_i(\psi)\right] \mid A_i, L_i\right), \end{aligned} \tag{17}$$

for ψ, where $d\{U_i(\psi), A_i, L_i\}$ is an arbitrary index function of the dimension of ψ; e.g. $d\{U_i(\psi), A_i, L_i\} = (1\ L_i)' A_i U_i(\psi)$. When A is dichotomous (coded 0/1), a locally efficient estimator of ψ^* (under the SDM together with the ignorability assumption (13)) can be obtained by solving (17) with $d\{U(\psi), A, L\} = E\{S_\psi(\psi) \mid U(\psi), A, L\}$, where $S_\psi(\psi)$ is the score for ψ under the observed data likelihood

$$\frac{\partial U(\psi^*)}{\partial Y} f(L) f\{U(\psi^*) \mid L\} f(A \mid L). \tag{18}$$

with all components substituted by suitable parametric models (Robins, 1997). For instance, under model (7) with $U(\psi^*)$ given L following a normal distribution, $S_\psi(\psi) = (1\ L)' A\{aU(\psi) + bL + c\}$ for certain constants a, b, c, so that a locally efficient estimator is obtained by solving (15).

Estimating equations of form (14) and (17) may also be used for repeated measures outcomes. In (14), $d^*(A_i, L_i)$ now becomes a $p \times (K+1)$-dimensional matrix, with p the dimension of ψ. In (17), $d\{U_i(\psi), A_i, L_i\}$ remains an arbitrary index function of the dimension of ψ; e.g. $d\{U_i(\psi), A_i, L_i\} = (1\ L_i)' A_i \sum_{m=1}^{K+1} U_{im}(\psi)$.

Remark. Note that the SMM together with assumption (13) is the same model for the observables as the semiparametric regression model (Chamberlain, 1987):

$$g\{E(Y \mid L, A)\} = \omega(L) + \gamma^*(L, A; \psi^*), \tag{19}$$

with $\omega(L)$ unspecified. Likewise, the SAFTM (with e.g. $\gamma(t, a, l; \psi) = t \exp(-a\psi)$) together with assumption (13) can be viewed as a semi-parametric generalisation of the accelerated failure time model (Wei, 1992), defined by $\log T = \psi A + \epsilon$ with $\epsilon \perp\!\!\!\perp A|L$.

Because of the curse of dimensionality, evaluating the conditional expectations appearing in equations (14) and (17) requires a parametric working model $\mathcal{A}$ for the conditional distribution of the exposure A:

$$f(A \mid L) = f(A \mid L; \alpha^*);$$

here $f(A \mid L; \alpha)$ is a known density function, smooth in α, and α^* is an unknown finite-dimensional parameter. For instance, for dichotomous exposure, one could assume that $P(A = 1 \mid L) = \text{expit}(\alpha_0^* + \alpha_1^* L)$ with $\alpha^* = (\alpha_0^*, \alpha_1^*)'$. Here, α^* can be estimated via standard (maximum likelihood) methods.

Evaluating (14) and (17) moreover requires a parametric working model $\mathcal{B}$ for the conditional distribution of $U(\psi^*)$ or the conditional expectation of $U^*(\psi^*)$. For (17), we model:

$$f\{U(\psi^*) \mid L\} = f\{U(\psi^*) \mid L; \beta^*\},$$

where $f\{U(\psi^*) \mid L; \beta\}$ is a known density function, smooth in β, and β^* is an unknown finite-dimensional parameter; to evaluate equations (14), specification of the conditional mean of $U^*(\psi^*)$, given L, suffices. For instance for a continuous outcome, one could assume that, conditional on L and for given ψ^*, $U(\psi^*) = Y - \psi_0^* A - \psi_1^* AL$ is normally distributed with mean $\beta_0^* + \beta_1^* L$ and variance β_2^{*2}, with $\beta^* = (\beta_0^*, \beta_1^*, \beta_2^*)'$. For each fixed value of ψ^*, β^* can be estimated using standard regression methods.

A consistent estimator of ψ^* indexing the SMM or SDM can now be obtained by solving equations (14) or (17), respectively, with α^* and β^* substituted by consistent estimators under models $\mathcal{A}$ and $\mathcal{B}$, respectively. The resulting estimator of ψ^* is called a G-estimator. In SDMs and linear or loglinear SMMs, it has the attractive property of being doubly robust (Robins and Rotnitzky, 2001): consistent when either model $\mathcal{A}$ or model $\mathcal{B}$ is correctly specified (in addition to a correctly specified structural model and ignorability); it does not require both to be correctly specified, nor does it require specifying which of both is correctly specified. That the solution to equation (14) is doubly robust can be seen because this equation has mean zero at $\psi = \psi^*$ when either model $\mathcal{A}$ or model $\mathcal{B}$ is correctly specified, even if one of them is misspecified. Equation (17) is likewise seen to have mean zero at $\psi = \psi^*$ under model $\mathcal{B}$; that it also has mean zero under model $\mathcal{A}$ at $\psi = \psi^*$ is seen by rewriting the equation as

$$\begin{aligned} 0 &= \sum_{i=1}^{n} d\left\{U_i(\psi), A_i, L_i\right\} - E\left[d\left\{U_i(\psi), A_i, L_i\right\} \mid L_i, A_i\right] \\ &\quad - E\left(d\left\{U_i(\psi), A_i, L_i\right\} - E\left[d\left\{U_i(\psi), A_i, L_i\right\} \mid L_i, A_i\right] \mid U_i(\psi), L_i\right). \end{aligned}$$

The result now follows, provided that the parameters α and β are variation-independent (i.e., not functionally related), so that a consistent estimator of α^* does not require consistent estimation of β^* and vice versa. Sandwich standard errors are obtained via the usual estimating equations theory.

In logistic SMMs, to the best of our knowledge, no estimators of ψ^* have been found that are root-n consistent under model $\mathcal{A}$ and the ignorability assumption. This is because the evaluation of $U^*(\psi)$ is anyway dependent upon a model for the conditional mean $E(Y \mid A, L)$ (see (3)). Tchetgen Tchetgen, Robins and Rotnitzky (2010) show that double robustness can instead be attained against misspecification of either a model for the density $f(Y \mid A = 0, L)$ or a model for the density $f(A \mid Y = 0, L)$. Their key to estimation of ψ^* is that the parameterized association $\gamma^*(L, A; \psi)$ between A and Y, when evaluated at $\psi = \psi^*$, can be used to render A and Y conditionally independent (given L) via inverse probability weighting. Their results apply equally to case-control designs (Tchetgen Tchetgen and Rotnitzky, 2011).

For Structural Mean Interaction Models, inference is developed in Vansteelandt et al. (2008a) when $g(.)$ is the identity or log link and in Tchetgen Tchetgen (2012) when $g(.)$ is the logistic link. Tchetgen Tchetgen and Robins (2010) focus on case-only designs and note that when $g(\cdot)$ is the log link, the multiplicative interaction (5) is identical to the conditional odds ratio between $A^{(1)}$ and $A^{(2)}$, given L within the subgroup of cases. This enables the use of results on logistic SMMs (Tchetgen Tchetgen, Robins and Rotnitzky, 2010) for robust estimation of multiplicative interactions under outcome-dependent sampling.

3.3 Censoring

Censoring presents additional challenges for the analysis of failure-time outcomes T. Random censoring or loss to follow-up can be dealt with through inverse probability of censoring weighting (Robins et al., 1992). Type I censoring, also known as censoring by end of follow-up, can be ignored in the analysis of SCFTMs, but must be dealt with in a different fashion in the analysis of SAFTMs. This is because $U(\psi^*)$ involves the failure-time itself, which is missing for all subjects who

fail after planned end-of-follow-up; the coarsening process is informative here as it depends on the actual failure time. We will next describe how Type I censoring can be dealt with in the analysis of SAFTMs.

Let C denote the planned end of follow-up time for given individual. C is known for all subjects, even those observed to fail. However, $U(\psi)$ cannot be evaluated for those who do not fail prior to time C. Knowing that $U(\psi^*) \perp\!\!\!\perp A \mid L$ under ignorability, we would then like to find a function $q\{U(\psi), C\}$ which is observable for all individuals and for which

$$q\{U(\psi^*), C\} \perp\!\!\!\perp A \mid L.$$

If such function is found, then ψ^* can be estimated by solving the original estimating equations for SDMs with $q\{U(\psi), C\}$ replacing $U(\psi)$. A natural choice would be $q\{U(\psi), C\} = \min\{U(\psi), U(C, A, L; \psi)\}$ with $U(C, A, L; \psi)$ the blipped-down censoring time, which is defined like $U(\psi)$ but with T substituted by C. However, this choice would not satisfy the required conditional independence property. The reason is that since C is fixed by design, $U(C, A, L; \psi)$ will in general be a function of A when $\psi \neq 0$ and so will generally fail to be conditionally independent of A, given L. Robins and Tsiatis (1991) thus propose to eliminate the dependence of $U(C, A, L; \psi)$ on A by redefining it to be $C(\psi) \equiv \min_a\{U(C, a, L; \psi)\}$. By thus minimizing over all feasible treatments a, any dependence on the observed treatment is broken so that $X(\psi) \equiv \min\{U(\psi), C(\psi)\}$ and $\Delta(\psi) \equiv I\{U(\psi) < C(\psi)\}$ become always observable quantities that are independent of A given L under ignorability, when evaluated at ψ^*. We may thus choose $q\{U(\psi), C\}$ to be an arbitrary function of $X(\psi)$ and $\Delta(\psi)$.

With each choice of $q\{U(\psi), C\}$, some subjects who are observed to fail may be treated as censored when $\psi \neq 0$. This can happen because for some subjects, $C(\psi)$ may be smaller than $U(\psi)$ even though $T < C$. Such subjects are called artificially censored. Artificial censoring has several consequences. Besides decreasing information about ψ^* as more subjects are artificially censored, the estimating equations are not, in general, continuous in ψ. This is because the functions $q\{U(\psi), C\}$ are not generally continuous in ψ, which happens in part because $\Delta(\psi)$ is not a smooth function of ψ. This can present problems for optimization, especially when ψ is a vector, and may moreover imply that the estimating equations have no solution in finite samples. This problem may be mitigated by choosing $q\{U(\psi), C\}$ to be a smooth function of ψ, e.g. $q\{U(\psi), C\} = \Delta(\psi) w_\alpha\{X(\psi)/C(\psi)\}$, where $w_\alpha(t) \equiv I(t > 1-\alpha)(1-t)/\alpha + I(t \leq 1-\alpha)$ (Joffe, Wang and Feldman, 2012). Vock et al. (2013) consider functions $q(\cdot; \psi)$ whose first derivatives exist for all ψ; they appear to have had better success in convergence for their optimization algorithm.

4. PROPERTIES OF G-ESTIMATION IN STRUCTURAL MODELS FOR POINT TREATMENTS UNDER IGNORABILITY

4.1 Comparison with ordinary regression estimators

Insight into the behaviour of G-estimators can be garnered by focusing on the simple model $\mathcal{M}_{\text{SMM}}$ defined by the ignorability assumption that $Y^a \perp\!\!\!\perp A|L$ for $a = 0, 1$, known treatment mechanism $f(A|L)$ and the SMM

$$E(Y^a - Y^0 \mid A = a, L) = \psi^* a.$$

Under homoscedasticity (i.e., when the conditional variance of the outcome, given A and L, is a constant σ^2), the locally efficient G-estimator of ψ^* under model $\mathcal{M}_{\text{SMM}}$ has influence function (Newey, 1990)

$$E\left\{\text{Var}(A \mid L)\right\}^{-1}\left\{A - E(A \mid L)\right\}\left\{Y - \psi^* A - E(Y - \psi^* A \mid L)\right\}; \tag{20}$$

it can thus in particular be obtained by setting the sample average of these influence functions to zero and solving for ψ^*. For binary treatment A, linear regression adjustment for the propensity score (Rosenbaum and Rubin, 1984) results in an estimator of ψ^* with influence function of the same form (20), but with $E(Y - \psi^* A \mid L)$ substituted by the population least squares fit from a regression of $Y - \psi^* A$ on the propensity score $E(A \mid L)$. Linear regression adjustment for the propensity score can therefore be viewed as an inefficient and non doubly robust G-estimation approach (Robins, Mark and Newey, 1992). The close relation between G-estimation and regression adjustment for the propensity score is not maintained in nonlinear models, where propensity score adjustment may not only demand correct models for the propensity score, but also for its association with outcome (Vansteelandt and Daniel, 2014). In nonlinear models, due to non-collapsibility of the treatment effect parameter (Greenland, Robins and Pearl, 1999), its meaning may also change depending on whether covariates are adjusted for in addition to the propensity score.

Ordinary regression estimators (in particular, maximum likelihood estimators obtained by fitting model (19) under a finite-dimensional parameterization of $\omega(L)$) are at least as efficient as the previously considered G-estimators, provided correct model specification. From the variance of the influence functions, we can deduce that the asymptotic variance of the locally efficient G-estimator is

$$\frac{\sigma^2}{E\left\{\text{Var}(A \mid L)\right\}}, \tag{21}$$

when there is homoscedasticity and the conditional mean $E(Y - \psi^* A \mid L) = E(Y|A = 0, L)$ is correctly specified. The ordinary least squares (OLS) estimator under the linear regression model $E(Y|A, L) = \beta' L + \psi A$ has an asymptotic variance which is smaller but, interestingly, not much smaller:

$$\frac{\sigma^2}{E\left[\text{Var}(A \mid L) + \left\{E(A \mid L) - \tilde{E}(A \mid L)\right\}^2\right]}.$$

This follows from its influence function, which is of the same form (20), but with $E(A \mid L)$ substituted by $\tilde{E}(A \mid L)$, the population least squares fit from a regression of A on L.

Despite their greater efficiency, ordinary regression estimators have a number of limitations not shared by G-estimators, an important one being their lack of extensibility to the analysis of sequential treatments (see Section 5). Furthermore, their explicit reliance on a model for the association between outcome and covariates can be disadvantageous when the treated and untreated subjects are very different in their covariate distributions, for then even well fitting models for the outcome may be prone to extrapolation bias (Rosenbaum and Rubin, 1984). This is not the case for G-estimators when they are based on a correctly specified

model ($\mathcal{A}$) for the treatment process. This is also seen from the form of the influence functions (20), following which individuals in regions of little or no overlap (i.e., at covariate values L where $\mathrm{Var}(A \mid L)$ is small) will hardly contribute in the calculation of the G-estimator because $A - E(A \mid L) \approx 0$ for such individuals. As with other estimation approaches based on propensity score adjustment (e.g. matching), the information about ψ^* will thus come primarily from regions with sufficient overlap, which we view as desirable. In contrast, OLS estimators are more susceptible to extrapolation bias since the leading term $A - \tilde{E}(A \mid L)$ in their influence functions may be far from zero for individuals in regions of little or no overlap. Finally, an advantage of G-estimation methods is that they can incorporate a priori knowledge on the exposure distribution. For instance, Vansteelandt et al. (2008b) exploit knowledge on the distribution of offspring genotypes given parental genotypes (based on Mendel's law of segregation), by using G-estimators to develop gene-environment interaction tests that are robust against misspecification of the effect of environmental exposures on the outcome.

4.2 Comparison with inverse probability weighted estimators

For the analysis of sequential treatments (see Section 5), marginal structural models (MSM) (Robins, Hernan and Brumback, 2000) and inverse probability weighted (IPW) estimators are much more popular than SMMs and SDMs and G-estimators. This is related to G-estimation being computationally more demanding by the lack of off-the-shelf software. It is thus of interest to compare the behaviour of these estimators in a simple setting with dichotomous treatment. Consider therefore model $\mathcal{M}_{\mathrm{MSM}}$, which is defined by the ignorability assumption that $Y^a \perp\!\!\!\perp A|L$ for $a = 0, 1$, known propensity score $E(A|L)$ and the nonparametric MSM

$$E(Y^a) = \alpha + \psi^* a.$$

Note, since $Y^a \perp\!\!\!\perp A|L$ for $a = 0, 1$, that $\psi^* = E(Y^1 - Y^0)$ in both models $\mathcal{M}_{\mathrm{SMM}}$ and $\mathcal{M}_{\mathrm{MSM}}$, and thus defines the same parameter. Nonetheless, model $\mathcal{M}_{\mathrm{MSM}}$ is less restrictive than model $\mathcal{M}_{\mathrm{SMM}}$ in that it does not postulate that the treatment effect is homogeneous (i.e., constant over levels of L). This explains why the asymptotic variance of the locally efficient IPW estimator under model $\mathcal{M}_{\mathrm{MSM}}$, which has influence function (Robins, Rotnitzky and Zhao, 1994)

$$\begin{aligned}&\frac{A\{Y - E(Y \mid A = 1, L)\}}{E(A \mid L)} - \frac{(1-A)\{Y - E(Y \mid A = 0, L)\}}{1 - E(A \mid L)}\\ &+E(Y \mid A = 1, L) - E(Y \mid A = 0, L) - \psi^*,\end{aligned}$$

is strictly larger than the variance of the locally efficient G-estimator (unless A and L are independent, as may be the case when A refers to a randomized treatment, in which case they are equally efficient). In particular, the asymptotic variance of the locally efficient IPW estimator equals

$$\sigma^2 E\left\{\frac{1}{\mathrm{Var}(A \mid L)}\right\}, \tag{22}$$

when the treatment effect is homogeneous. The difference between (21) and (22) can be sizeable when the propensity score is close to zero or 1 for some values of L for then $\mathrm{Var}(A \mid L)$ is close to zero and thus $1/\mathrm{Var}(A \mid L)$ can take on large

values. In our opinion, this difference is not usually offset by the weaker restrictions imposed by the MSM. Indeed, the marginal treatment effect would seldom be of scientific interest when certain subjects are almost precluded from receiving treatment or no treatment. Moreover, the G-estimator retains a useful interpretation even when the assumption of constant treatment effects fails in the sense that $E(Y^1 - Y^0 \mid A = 1, L) = \psi(L)$ for some function $\psi(L)$. Indeed, in that case the locally efficient G-estimator converges to

$$\frac{E\left\{\mathrm{Var}(A \mid L)\psi(L)\right\}}{E\left\{\mathrm{Var}(A \mid L)\right\}}, \tag{23}$$

which continues to be useful as a weighted average of treatment effects $\psi(L)$, with most weight given to strata with most information about the treatment effect.

This difference in asymptotic variance between both estimators becomes even more pronounced in the likely event that the model for $E(Y \mid A = 0, L) = E(Y^0 \mid L) = E(Y - \psi^* A \mid L)$ is misspecified. Let $\Delta(L) = E(Y \mid A = 0, L) - E^*(Y \mid A = 0, L)$ denote the degree of misspecification at covariate value L, with $E(Y \mid A = 0, L)$ the true expectation and $E^*(Y \mid A = 0, L)$ the expectation used for evaluating the locally efficient G-estimator. Furthermore, assume that in truth the treatment effect is homogeneous. Then the the asymptotic variance of the G-estimator becomes

$$\frac{\sigma^2}{E\left\{\mathrm{Var}(A \mid L)\right\}} + \frac{E\left\{\mathrm{Var}(A \mid L)\Delta(L)^2\right\}}{E\left\{\mathrm{Var}(A \mid L)\right\}^2},$$

and the asymptotic variance of the previously considered IPW estimator becomes

$$\sigma^2 E\left\{\frac{1}{\mathrm{Var}(A \mid L)}\right\} + E\left[\frac{\left\{\Delta(L) + \psi^* E(1 - A \mid L)\right\}^2}{\mathrm{Var}(A \mid L)}\right].$$

Consider now that model misspecification is more likely in regions of little overlap. Then because $\mathrm{Var}(A \mid L) \approx 0$ in these regions, model misspecification in these regions will only have a minor impact on the variance of the G-estimator, but a particularly strong impact on the variance of the locally efficient IPW estimator. Similar findings have been noted concerning the finite-sample bias of these estimators (Vansteelandt, Bekaert and Claeskens, 2012).

While this contrast between G-estimation and IPW-estimation may potentially be less dramatic when the propensity score is not considered as fixed and known, it more likely understates the factual differences if one considers that mainstream applications are based on sequential treatments (and thus even more variable inverse probability weights) and on simple, inefficient inverse probability weighting methods. The latter can be viewed as inducing extreme misspecification in the outcome model as they amount to setting $E(Y \mid A = 1, L) = E(Y \mid A = 0, L) = 0$. We thus believe that more routine application of G-estimation is warranted.

5. STRUCTURAL NESTED MODELS FOR TIME-VARYING TREATMENTS

Before introducing SNMs for time-varying or sequential treatments, we consider the structure of observed data in observational studies with repeated treatments and covariates, as well as definitions of causal effects in such setting. Suppose that measurements are collected at fixed time points $t_0, t_1, ..., t_{K+1}$. Let A_k

denote the treatment provided at time $t_k, k = 0, ..., K$, and L_k denote other covariates measured at that time; Y_k, the outcome measured at time $t_k, k = 1, ..., K+1$, is part of L_k. We presume the variables are ordered L_0, A_0, L_1, A_1, etc.; thus, covariates and outcome at t_k precede treatment at t_k.

Let $Y_m^{\overline{a}_{m-1}}$ denote the outcome that would be seen at time t_m in a given individual were (s)he to receive treatment history $\overline{a}_{m-1}$ through time t_{m-1}. The variables $Y_m^{\overline{a}_{m-1}}$ are potential outcomes, which are again linked to the observed data via the consistency assumption that $Y_m = Y_m^{\overline{a}_{m-1}}$ if $\overline{A}_{m-1} = \overline{a}_{m-1}$. We presume that treatment at or after t_m cannot affect outcome at times up to t_m; thus, $\underline{Y}_m^{\overline{a}_{m-1},\underline{a}_m} = \underline{Y}_m^{\overline{a}_{m-1},\underline{a}_m^\dagger}$ for $\underline{a}_m \neq \underline{a}_m^\dagger$. Causal effects can now be defined as comparisons of potential outcomes $\underline{Y}^{\overline{a}_K}$ for the same group of subjects for different treatment histories $\overline{a}_K$, $\overline{a}_K^\dagger$, $\overline{a}_K \neq \overline{a}_K^\dagger$ (Robins, 1986). If the outcome is measured only at the end of a fixed follow-up period, or only at a subset of the follow-up times, we can let $Y_m = \cdot$, where $\cdot$ denotes missing or undefined values for the times where the outcome is not measured. Most of the subsequent presentation then applies to those settings.

5.1 Structural nested mean models

Structural nested mean models (SNMMs) (Robins, 1994; Robins, Rotnitzky and Scharfstein, 2000) simulate the sequential removal of an amount (blip) of treatment at t_m on subsequent average outcomes, after having removed the effects of all subsequent treatments. Given a history $\overline{a}_m$, define the counterfactual history $(\overline{a}_m, 0)$ as the history $\overline{a}^\dagger$ that agrees with $\overline{a}_m$ through time t_m and is 0 thereafter. SNMMs then model the effect of a blip of treatment at t_m on the subsequent outcome means when holding all future treatments fixed at their reference level 0; thus, they parameterize contrasts of $\underline{Y}_{m+1}^{\overline{a}_m,0}$ and $\underline{Y}_{m+1}^{\overline{a}_{m-1},0}$ conditionally on treatment and covariate histories through t_m as

$$\begin{aligned} &g\left\{E\left(\underline{Y}_{m+1}^{\overline{a}_m,0} \mid \overline{L}_m = \overline{l}_m, \overline{A}_m = \overline{a}_m\right)\right\} \\ &-g\left\{E\left(\underline{Y}_{m+1}^{\overline{a}_{m-1},0} \mid \overline{L}_m = \overline{l}_m, \overline{A}_m = \overline{a}_m\right)\right\} = \gamma_m^*(\overline{l}_m, \overline{a}_m; \psi^*), \end{aligned}$$

for each $m = 0, ..., K$ and $(\overline{l}_m, \overline{a}_m)$, where $\gamma_m^*(\overline{l}_m, \overline{a}_m; \psi)$ is a known $(K+1-m)$-dimensional function, smooth in ψ, and for each $\overline{l}_m, \overline{a}_{m-1}$ and ψ it is by definition required that $\gamma_m^*(\overline{l}_m, \overline{a}_{m-1}, 0; \psi) = 0$. Alternatively, one may focus on the effect of treatment on the end-of-study outcome $Y \equiv Y_{K+1}$ only, in which case one obtains a SNMM of the form

$$\begin{aligned} &g\left\{E\left(Y^{\overline{a}_m,0} \mid \overline{L}_m = \overline{l}_m, \overline{A}_m = \overline{a}_m\right)\right\} \\ &-g\left\{E\left(Y^{\overline{a}_{m-1},0} \mid \overline{L}_m = \overline{l}_m, \overline{A}_m = \overline{a}_m\right)\right\} = \gamma_m^*(\overline{l}_m, \overline{a}_m; \psi^*), \end{aligned}$$

for each $m = 0, ..., K$ and $(\overline{l}_m, \overline{a}_m)$, where $\gamma_m^*(\overline{l}_m, \overline{a}_m; \psi)$ is now 1-dimensional. The above contrasts generalize the notion of the effect of treatment on the treated to the setting of a sequence of treatments. The name 'nested' refers to the nesting across time, of the subgroups defined by $\overline{L}_m$ and $\overline{A}_m$ within which the effects are evaluated.

Typically, the parameterization is chosen to be such that $\gamma_m^*(\overline{l}_m, \overline{a}_m; 0) = 0$ for all $\overline{l}_m, \overline{a}_m$ so that $\psi = 0$ encodes the null hypothesis of no treatment effect. For

instance, with 2 time points ($K = 1$) a linear SNMM may be given by:

$$\begin{aligned} E(Y_2^{(a_0,a_1)} - Y_2^{(a_0,0)} \mid \overline{L}_1 = \bar{l}_1, \overline{A}_1 = \overline{a}_1) &= (\psi_0^* + \psi_1^* l_1 + \psi_2^* a_0) a_1 \\ E(Y_2^{(a_0,0)} - Y_2^0 \mid L_0 = l_0, A_0 = a_0) &= (\psi_3^* + \psi_4^* l_0) a_0 \\ E(Y_1^{(a_0,0)} - Y_1^0 \mid L_0 = l_0, A_0 = a_0) &= (\psi_0^* + \psi_1^* l_0) a_0. \end{aligned}$$

Here, the first equation models the effect of A_1 on Y_2, the second models the effect of A_0 on Y_2 and the third models the effect of A_0 on Y_1, all within levels of variables defined prior to the considered exposure. Thus ψ_0^*, ψ_1^* and ψ_2^* encode short-term treatment effects, which are here assumed to be constant at all time points, and ψ_3^* and ψ_4^* encode long-term treatment effects. These effects are visualised in Figures 3 and 4 below. When interest merely lies in the effect on the end-of-study outcome, then the above model for Y_1 can be ignored.

Figure 2 about here: 'Visualisation of the effects $E(Y_1^{(a_0,0)} - Y_1^0 \mid L_0 = l_0, A_0 = a_0)$ and $E(Y_2^{(a_0,a_1)} - Y_2^{(a_0,0)} \mid \overline{L}_1 = \bar{l}_1, \overline{A}_1 = \overline{a}_1)$. Lines within the circles depict covariate strata; lines outside the circles depict exposure strata.'

Figure 3 about here: 'Visualisation of the effects $E(Y_2^{(a_0,0)} - Y_2^0 \mid L_0 = l_0, A_0 = a_0)$. Lines within the circles depict covariate strata; lines outside the circles depict exposure strata.'

Under the SNMM, as in Section 2.1, it is possible to define a transformation $U_m^*(\psi^*)$ of $\underline{Y}_{m+1}$, whose mean value equals the mean that would be observed if treatment were suspended from time t_m onwards, in the sense that

$$E\left\{U_m^*(\psi^*) \mid \overline{L}_m, \overline{A}_m\right\} = E\left(\underline{Y}_{m+1}^{\overline{a}_{m-1},0} \mid \overline{L}_m, \overline{A}_m\right), \tag{24}$$

for $m = 0, ..., K$. Here, $U_m^*(\psi)$ is a vector with components

$$Y_k - \sum_{l=m}^{k-1} \gamma_{l,k}^*(\overline{L}_l, \overline{A}_l; \psi),$$

for $k = m+1, ..., K+1$ (or for $k = K+1$ only if interest merely lies in the effect on the end-of-study outcome) if $g(.)$ is the identity link, and

$$Y_k \exp\left\{-\sum_{l=m}^{k-1} \gamma_{l,k}^*(\overline{L}_l, \overline{A}_l; \psi)\right\},$$

if $g(.)$ is the log link. These equations formalise the notion of peeling off or blipping down the treatment effects over the treatment period from t_m to t_{k-1}. For instance, in the previous example for 2 time points,

$$\begin{aligned} U_1^*(\psi) &= Y_2 - (\psi_0^* + \psi_1^* L_1 + \psi_2^* A_0) A_1 \\ U_0^*(\psi) &= (Y_1 - (\psi_0^* + \psi_1^* L_0) A_0, Y_2 - (\psi_0^* + \psi_1^* L_1 + \psi_2^* A_0) A_1 - (\psi_3^* + \psi_4^* L_0) A_0)'. \end{aligned}$$

For link functions other than the identity and log link, such a transformation can still be defined, but depends on the observed data distribution in a complicated

and contrived way. For instance, when $g(.)$ is the logit link and there are 2 time points ($K = 1$), then under the SNMM we have that

$$\begin{aligned} E\left(Y_2^0|L_0 = l_0, A_0 = a_0\right) &= g^{-1}\left[g\left\{E\left(g^{-1}\left[g\left\{E\left(Y_2|\overline{L}_1, A_1, A_0 = a_0\right)\right\}\right.\right.\right.\right. \\ &\quad \left.\left.\left.\left. -\gamma_1^*(\overline{L}_1, A_1, A_0 = a_0; \psi^*)\right] |L_0 = l_0, A_0 = a_0\right)\right\} - \gamma_0^*(l_0, a_0; \psi^*)\right]. \end{aligned}$$

The calculation of $U_0(\psi)$ thus not only demands knowledge of $E\left(Y_2|\overline{L}_1, \overline{A}_1\right)$, but also of the distribution of (L_1, A_1), given (L_0, A_0).

The effect of a sequential treatment on the failure time distribution can be parameterised through a collection of SNMMs with log link, one for each time point (Robins and Hernan, 2009; Picciotto et al., 2011). In continuous time (Martinussen et al., 2011), such structural nested cumulative failure time models are defined by restrictions of the form:

$$\frac{P\left(T^{\overline{a}_m,0} > t \mid \overline{L}_m = \overline{l}_m, \overline{A}_m = \overline{a}_m, T \geq t_m\right)}{P\left(T^{\overline{a}_{m-1},0} > t \mid \overline{L}_m = \overline{l}_m, \overline{A}_m = \overline{a}_m, T \geq t_m\right)} = \exp\left\{\gamma_m^*(t, \overline{l}_m, \overline{a}_m; \psi^*)\right\},$$

for all t and $m = 0, ..., K$, where $\gamma_m^*(t, \overline{l}_m, \overline{a}_m; \psi)$ is a known function, smooth in ψ and monotonic in t, and $\gamma_m^*(t, \overline{l}_m, \overline{a}_{m-1}, 0; \psi) = 0$ for all $t, \overline{l}_m, \overline{a}_{m-1}$ and ψ.

5.2 Structural nested distribution models

Structural nested distribution models (SNDMs) are closely related to SNMMs, but parameterise a map between percentiles of the distribution of $Y_k^{\overline{a}_m,0}$ and percentiles of the distribution of $Y_k^{\overline{a}_{m-1},0}$. They are most easily understood by first considering the class of more restrictive rank-preserving SNDMs. In particular, for each exposure $A_m, m = 0, ..., K$, let us first consider a rank-preserving SNDM to parameterise its effect on the end-of-study outcome Y.

$$Y^{\overline{a}_{m-1},0} = \gamma_m\left(Y^{\overline{a}_m,0}, \overline{l}_m, \overline{a}_m; \psi^*\right),$$

for subjects with $\overline{A}_m = \overline{a}_m$ and $\overline{L}_m = \overline{l}_m$, $m = 0, ..., K$. Here, $\gamma_m(y, \overline{l}_m, \overline{a}_m; \psi)$ is a known function, smooth in ψ and a smooth, monotonic function of y, which contrasts the counterfactuals $Y^{\overline{a}_{m-1},0}$ and $Y^{\overline{a}_m,0}$, and must satisfy $\gamma_m(y, \overline{l}_m, \overline{a}_{m-1}, 0; \psi) = y$ for all y and ψ. For instance, with 2 time points ($K = 1$) a rank preserving SNDM may be given by the following set of restrictions:

$$\begin{aligned} Y^{A_0,0} &= \gamma_1(Y, \overline{L}_1, \overline{A}_1) = Y - (\psi_1^* + \psi_2^* L_1)A_1 \\ Y^0 &= \gamma_0(Y^{A_0,0}, L_0, A_0) = Y^{A_0,0} - (\psi_1^* + \psi_2^* L_0)A_0 \\ &= Y - \psi_1^*(A_0 + A_1) - \psi_2^*(L_1 A_1 + L_0 A_0). \end{aligned}$$

A SNDM relaxes these restrictions by demanding that they merely hold in distribution, conditional on the observed history (i.e. $\overline{L}_m = \overline{l}_m$ and $\overline{A}_m = \overline{a}_m$).

To describe the effect on a repeated counterfactual future $\underline{Y}_{m+1}^{\overline{a}_m,0}$, we can borrow ideas from Section 2.3.2. In particular, upon substituting A by A_m, L by $\overline{L}_m, \overline{A}_{m-1}$ and Y_k by $Y_{m+k}^{\overline{a}_m,0}$ in the rank-preserving model (9), we obtain the identity:

$$Y_{m+k}^{\overline{a}_{m-1},0} = \gamma_{m,m+k}\left(Y_{m+1:m+k}^{\overline{a}_m,0}, \overline{l}_m, \overline{a}_m; \psi^*\right), \tag{25}$$

for subjects with $\overline{A}_m = \overline{a}_m$ and $\overline{L}_m = \overline{l}_m$, $m = 0, ..., K$ and $k = 1, ..., K+1-m$. Here, $\gamma_{m,m+k}(y_{m:m+k}, \overline{l}_m, \overline{a}_m; \psi)$ is a known function, smooth in ψ and a smooth, monotonic function of y_{m+k}, which contrasts the counterfactuals $Y_{m+k}^{\overline{a}_{m-1},0}$ and $Y_{m+k}^{\overline{a}_m,0}$, and must satisfy $\gamma_{m,m+k}(y_{m:m+k}, \overline{l}_m, \overline{a}_{m-1}, 0; \psi) = y_{m+k}$ for all $y_{m+k}, \overline{l}_m, \overline{a}_{m-1}$ and ψ. For instance, with 2 time points ($K = 1$) a rank preserving SNDM may be given by the following set of restrictions:

$$
\begin{aligned}
(26) \qquad Y_1^0 &= \gamma_{0,1}(Y_1, L_0, A_0) = Y_1 - (\psi_1^* + \psi_2^* L_0)A_0 \\
Y_2^{A_0,0} &= \gamma_{1,2}(Y_2, \overline{L}_1, \overline{A}_1) = Y_2 - (\psi_1^* + \psi_2^* L_1)A_1 \\
Y_2^0 &= \gamma_{0,2}(Y_1, Y_2^{A_0,0}, L_0, A_0) = Y_2^{(A_0,0)} - (\psi_3^* + \psi_4^* Y_1)A_0 \\
(27) \qquad &= Y_2 - (\psi_1^* + \psi_2^* L_1)A_1 - (\psi_3^* + \psi_4^* Y_1)A_0.
\end{aligned}
$$

Here, the first two equations express short-term exposure effects, i.e., the effect of A_0 on Y_1 and of A_1 on Y_2. The third equation expresses the effect of A_0 on Y_2 (more precisely, its effect on $Y_2^{A_0,0}$). As in Section 2.3.2, this equation must take into account that the effect may be different depending on the outcome level at time t_1; this allows for A_0 to also affect the dependence between Y_1 and Y_2, but evidently complicates interpretation. More generally, rank-preserving SNDMs allow for the effect of a_m on Y_{m+k}, as encoded by a contrast of $Y_{m+k}^{\overline{a}_m,0}$ and $Y_{m+k}^{\overline{a}_{m-1},0}$, to depend on the history of treatments and covariates up to time t_m, but additionally on the potential outcome history under the treatment regime $(\overline{a}_m, 0)$, up to time t_{m+k-1}.

A SNDM relaxes the restrictions of a rank preserving SNDM by demanding that the equality (25) merely holds in distribution, conditional on $\overline{L}_m = \overline{l}_m$ and $\overline{A}_m = \overline{a}_m$. Assuming that for given $(\overline{L}_m, \overline{A}_m)$, $\underline{Y}_{m+1}$ has a continuous multivariate distribution with probability 1, a SNDM can thus be defined by:

(28)
$$
F_{\underline{Y}_{m+1}^{\overline{a}_{m-1},0} \mid \overline{L}_m = \overline{l}_m, \overline{A}_m = \overline{a}_m} \left\{ \gamma_m(\underline{y}_{m+1}, \overline{l}_m, \overline{a}_m; \psi^*) \right\} = F_{\underline{Y}_{m+1}^{\overline{a}_m,0} \mid \overline{L}_m = \overline{l}_m, \overline{A}_m = \overline{a}_m} \left(\underline{y}_{m+1} \right),
$$

for all $\overline{l}_m, \overline{a}_m$, where $\gamma_m(\underline{y}_{m+1}, \overline{l}_m, \overline{a}_m; \psi^*)$ is a vector with components $\gamma_{m,k}(y_{m+1:m+k}, \overline{l}_m, \overline{a}_m; \psi^*)$ for $k = 1, ..., K+1-m$, where the components $\gamma_{m,k}$ are defined in recursive fashion similar to in Section 2.3.2.

Under the SNDM, a variable $U_m(\psi^*) = (U_{m,m+1}(\psi^*), ..., U_{m,K+1}(\psi^*))'$ can be constructed which predicts how the outcomes past time t_m would look like if treatment were suspended from time t_m onwards, in the sense that:

(29)
$$
P\left\{ U_m(\psi^*) > \underline{y}_{m+1} \mid \overline{L}_m, \overline{A}_m = \overline{a}_m \right\} = P\left(\underline{Y}_{m+1}^{\overline{a}_{m-1},0} > \underline{y}_{m+1} \mid \overline{L}_m, \overline{A}_m = \overline{a}_m \right).
$$

This variable can be recursively obtained for $m = K, ..., 0$ from

$$
(30) \qquad U_{m,m+k}(\psi) \equiv \gamma_{m,m+k}\left\{ (Y_{m+1}, U_{m+1:m+k}(\psi)), \overline{L}_m, \overline{A}_m; \psi \right\},
$$

for $k = 1, ..., K+1-m$. For instance, in the SNDM that assumes the identities in (26) hold in distribution (conditional on the observed history), we have that

$$
\begin{aligned}
U_1(\psi) &= Y_2 - (\psi_1 + \psi_2 L_1)A_1 \\
U_0(\psi) &= (Y_1 - (\psi_1 + \psi_2 L_0)A_0, Y_2 - (\psi_1 + \psi_2 L_1)A_1 - (\psi_3 + \psi_4 Y_1)A_0).
\end{aligned}
$$

The identity (30) will be useful in estimation and for predicting the effect of specific interventions on the outcome distribution.

Structural nested failure time models (SNFTMs) are a variant of SNDMs which have seen most applications to date. These link percentiles from the conditional distributions of $T^{\overline{a}_{m-1},0}$ and $T^{\overline{a}_m,0}$, conditional on $\overline{L}_m, \overline{A}_m = \overline{a}_m$, and for subjects who are still in the risk set (say, alive) at time t_m:

$$S_{T^{\overline{a}_{m-1},0}|\overline{L}_m=\overline{l}_m,\overline{A}_m=\overline{a}_m,T\geq t_m}\left\{\gamma_m(t,\overline{l}_m,\overline{a}_m;\psi^*)\right\} = S_{T^{\overline{a}_m,0}|\overline{L}_m=\overline{l}_m,\overline{A}_m=\overline{a}_m,T\geq t_m}(t),$$

for $t > t_m$, where $S(\cdot)$ denotes a survival function. Here, $\gamma_m(t,\overline{l}_m,\overline{a}_m;\psi^*)$ is a known function, smooth in ψ and monotonic in t, and $\gamma_m(t,\overline{l}_m,\overline{a}_{m-1},0;\psi) = t$ for all $t, \overline{l}_m, \overline{a}_{m-1}$ and ψ. For instance, the choice $\gamma_m(t,\overline{l}_m,\overline{a}_m;\psi) = t_m + \min(1, t - t_m)\exp(a_m\psi)$ expresses that the effect of suspending treatment a_m at time t_m is to shorten the residual lifetime $t - t_m$ with a factor $\exp(a_m\psi)$. For this choice of model, one can predict amongst individuals who survive to (or through, or until) time t_m what their lifetime would be had treatment been suspended from time t_m onwards, as

$$U_m(\psi) \quad = \quad t_m + \sum_{k:t_m\leq t_k\leq T} \exp\left(A_k\psi\right) + (T - t_{T^-})\exp(A_{t_{T^-}}\psi),$$

where t_{T^-} denotes the largest time point in $\{t_0, ..., t_K\}$ less than T and $U_m(\psi)$ is a random variable for which

$$P\left\{U_m(\psi^*) > t \mid \overline{L}_m, \overline{A}_m = \overline{a}_m, T \geq t_m\right\} = P\left(T^{\overline{a}_{m-1},0} > t \mid \overline{L}_m, \overline{A}_m = \overline{a}_m, T \geq t_m\right).$$

5.3 Retrospective blip models

Retrospective blip models have been extended to model the effect of a sequential treatment on a scalar end-of-study outcome $Y \equiv Y_{K+1}$ conditional on the treatment and covariate history up to end-of-study. Mean models take the form

$$\begin{aligned} & g\left\{E\left(Y^{\overline{a}_m,0}|\overline{L}_K = \overline{l}_K, \overline{A}_K = \overline{a}_K\right)\right\} \\ & -g\left\{E\left(Y^{\overline{a}_{m-1},0}|\overline{L}_K = \overline{l}_K, \overline{A}_K = \overline{a}_K\right)\right\} = \gamma^*_m(\overline{l}_K, \overline{a}_K; \psi^*), \end{aligned} \tag{31}$$

where $\gamma^*_m(\overline{l}_K, \overline{a}_K; \psi)$ is a known function, smooth in ψ and equalling zero for all $\psi, \overline{l}_K$ and $\overline{a}_K$ with $a_m = 0$. Distribution models take the form:

$$F_{Y^{\overline{a}_{m-1},0}|\overline{L}_K=\overline{l}_K,\overline{A}_K=\overline{a}_K}\left\{\gamma_m(y,\overline{l}_K,\overline{a}_K;\psi^*)\right\} = F_{Y^{\overline{a}_m,0}|\overline{L}_K=\overline{l}_K,\overline{A}_K=\overline{a}_K}(y),$$

where $\gamma_m(y, \overline{l}_K, \overline{a}_K; \psi)$ is a known function, smooth in ψ and equalling y for all $\psi, y, \overline{l}_K$ and $\overline{a}_K$ with $a_m = 0$; a rank-preserving version of this was proposed by Joffe, Small And Hsu (2007). For nonparametric identifiability, restrictions are needed on the functions $\gamma^*_m(\overline{l}_K, \overline{a}_K; \psi^*)$ and $\gamma_m(y, \overline{l}_K, \overline{a}_K; \psi)$, e.g. that they do not involve the future $\underline{a}_{m+1}$ and $\underline{l}_{m+1}$ (Vansteelandt, 2010).

Retrospective blip models can be useful for modelling a dichotomous outcome (Vansteelandt, 2010). Under these models, identity (24) is satisfied with $U^*_m(\psi)$ being a vector with components

$$g^{-1}\left[g\left\{E(Y|\overline{L}_K, \overline{A}_K)\right\} - \sum_{l=m}^{k-1} \gamma^*_{l,k}(\overline{L}_K, \overline{A}_K; \psi)\right].$$

Evaluation of $U_m^*(\psi)$ (which is needed to make estimation of ψ^* manageable) then merely requires a model for $E(Y|\overline{L}_K, \overline{A}_K)$, but not for the distribution of treatment and covariates at each time. The parameters indexing these models are nonetheless more limited than the parameters indexing SNMMs in that they cannot be used by themselves for making treatment decisions prior to the end-of-study time, unless one integrates over the distribution of covariates subsequent to m (see e.g. Vansteelandt, 2010).

6. IDENTIFICATION AND ESTIMATION IN STRUCTURAL NESTED MODELS FOR SEQUENTIAL TREATMENTS

This section sketches identifying assumptions and inferential methods for sequential treatments. Under instrumental variables assumptions sketched in Section 6.3 and under the future ignorability assumptions sketched in Section 6.2, inferential methods have been developed for SNMs, but these assumptions do not suffice for the identification of marginal treatment effects and hence parameters indexing MSMs. The broader array of useful identifying assumptions thus constitutes an important advantage of SNMs.

6.1 Sequential ignorability

The assumption of ignorable treatment assignment can be generalised to sequential treatments as follows:

$$A_m \perp\!\!\!\perp \underline{Y}_{m+1}^{\overline{a}_{m-1},0} \mid \overline{L}_m, \overline{A}_{m-1} = \overline{a}_{m-1}, \tag{32}$$

for $m = 0, ..., K$. This assumption has been called variously 'no unmeasured confounders assumption', 'sequential ignorability', 'sequential randomization' or 'exchangeability'. It expresses that at each time t_m, the observed history of covariates $\overline{L}_m$ and exposures $\overline{A}_{m-1}$ includes all risk factors of A_m that are also associated with future outcomes.

This assumption together with identity (24) imply that

$$E\left\{U_m(\psi^*) \mid \overline{L}_m, \overline{A}_m\right\} = E\left\{U_m(\psi^*) \mid \overline{L}_m, \overline{A}_{m-1}\right\}$$

for all m under a SNMM. The parameter ψ^* indexing a SNMM can therefore be estimated by solving

$$\begin{aligned} 0 &= \sum_{i=1}^{n}\sum_{m=0}^{K}\left[d_m(\overline{L}_{im}, \overline{A}_{im}) - E\left\{d_m(\overline{L}_{im}, \overline{A}_{im}) \mid \overline{L}_{im}, \overline{A}_{i,m-1}\right\}\right] \\ &\quad \times \left[U_{im}(\psi) - E\left\{U_{im}(\psi) \mid \overline{L}_{im}, \overline{A}_{i,m-1}\right\}\right], \end{aligned} \tag{33}$$

where $d_m(\overline{L}_{im}, \overline{A}_{im}), m = 0, ..., K$ is an arbitrary $p \times (K + 1 - m)$-dimensional function, with p the dimension of ψ. This estimating equation essentially sets the sum across time points m of the conditional covariances between $U_{im}(\psi)$ and the given function $d_m(\overline{L}_{im}, \overline{A}_{im})$, given $\overline{L}_{im}, \overline{A}_{i,m-1}$, to zero. When the previous outcome is included in the confounder history (i.e. $\overline{L}_{im}$ includes Y_{im}) and there is homoscedasticity (i.e., when the conditional variance of $U_{im}(\psi^*)$ given $\overline{L}_{im}, \overline{A}_{im}$ is constant for $m = 0, ..., K$), then local semi-parametric efficiency under the SNMM is attained upon choosing

$$d_m(\overline{L}_{im}, \overline{A}_{im}) = E\left\{\frac{\partial U_m(\psi^*)}{\partial \psi} \mid \overline{L}_{im}, \overline{A}_{im}\right\}.$$

Sequential ignorability (32) together with identity (29) moreover implies that

$$U_m(\psi^*) \perp\!\!\!\perp A_m \mid \overline{L}_m, \overline{A}_{m-1}$$

for all m under the SNDM. This conditional independence restriction suggests that the parameter indexing a SNDM can be solved from

$$\begin{aligned} 0 &= \sum_{i=1}^{n}\sum_{m=0}^{K} d_m\left\{U_{im}(\psi), \overline{A}_{im}, \overline{L}_{im}\right\} - E\left[d_m\left\{U_{im}(\psi), \overline{A}_{im}, \overline{L}_{im}\right\} \mid \overline{L}_{im}, \overline{A}_{im}\right] \\ &\quad -E\left(d_m\left\{U_{im}(\psi), \overline{A}_{im}, \overline{L}_{im}\right\} - E\left[d_m\left\{U_{im}(\psi), \overline{A}_{im}, \overline{L}_{im}\right\} \mid \overline{L}_{im}, \overline{A}_{im}\right]\right. \\ &\quad \left. \mid U_{im}(\psi), \overline{L}_{im}, \overline{A}_{i,m-1}\right), \end{aligned} \tag{34}$$

where the index functions $d_m\left\{U_{im}(\psi), \overline{A}_{im}, \overline{L}_{im}\right\}$ must be of the dimension of ψ. When the previous outcome is included in the confounder history (i.e. $\overline{L}_{im}$ includes Y_{im}), then local semi-parametric efficiency is obtained upon choosing

$$d_m\left\{U_{im}(\psi), \overline{A}_{im}, \overline{L}_{im}\right\} = E\left\{S_\psi(\psi) \mid U_{im}(\psi), \overline{A}_{im}, \overline{L}_{im}\right\},$$

where $S_\psi(\psi)$ is the score for ψ under the observed data likelihood

$$\begin{aligned} f(\overline{Y}_{K+1}, \overline{L}_K, \overline{A}_K) &= f\left\{U_0(\psi^*)\right\} \prod_{m=0}^{K}\left[f\left\{L_m|\overline{L}_{m-1}, \overline{A}_{m-1}, U_m(\psi^*)\right\}\right. \\ &\quad \left. \times f\left\{A_m|\overline{L}_m, \overline{A}_{m-1}, U_m(\psi^*)\right\}\left|\frac{\partial U_m(\psi^*)}{\partial U_{m+1}(\psi^*)}\right|\right], \end{aligned}$$

with all components substituted by suitable parametric models (Robins, 1997); here, the term $f\left\{A_m|\overline{L}_m, \overline{A}_{m-1}, U_m(\psi^*)\right\} = f\left(A_m|\overline{L}_m, \overline{A}_{m-1}\right)$ under sequential igorability, and thus can be ignored. This likelihood formulation is of interest in itself because it enables specifying the joint distribution of the variables in a way that is consistent with the sharp null hypothesis of no effect under the assumption of sequential ignorability, even in the presence of confounding by variables affected by treatment.

Solving estimating equations (33) and (34) requires a parametric model $\mathcal{A}$ for the conditional distribution of the exposure A_m for $m = 0, ..., K$:

$$f(A_m \mid \overline{L}_{m-1}, \overline{A}_{m-1}) = f(A_m \mid \overline{L}_{m-1}, \overline{A}_{m-1}; \alpha^*),$$

where $f(A_m \mid \overline{L}_{m-1}, \overline{A}_{m-1}; \alpha)$ is a known density function, smooth in α, and α^* is an unknown finite-dimensional parameter which can be estimated via standard maximum likelihood. In addition, it requires a parametric model $\mathcal{B}$ for the conditional mean (or distribution) of $U_m^*(\psi^*)$ (or $U_m(\psi^*)$) for $m = 0, ..., K$:

$$f\left\{U_m(\psi^*) \mid \overline{L}_m, \overline{A}_{m-1}\right\} = f\left\{U_m(\psi^*) \mid \overline{L}_m, \overline{A}_{m-1}; \gamma^*\right\},$$

where $f\left\{U_m(\psi^*) \mid \overline{L}_m, \overline{A}_{m-1}; \gamma\right\}$ is a known density function, smooth in γ and γ^* is an unknown finite-dimensional parameter. As before, when the parameters α and γ are variation-independent, then so-called G-estimators that solve (33) and (34), obtained upon substituting α^* and γ^* by consistent estimators, are

doubly robust (Robins and Rotnitzky, 2001): consistent when the SNM and either model $\mathcal{A}$ or model $\mathcal{B}$ is correctly specified, regardless of which. This double robustness property of the G-estimator is desirable for various reasons. First, it provides justification for using simple models for the multivariate distribution $f\left\{U_m(\psi^*) \mid \overline{L}_m, \overline{A}_{m-1}\right\}$ or even setting $E\left\{U_{im}(\psi) \mid \overline{L}_{im}, \overline{A}_{i,m-1}\right\} = 0$ in (33). Second, while alternative proposals that rely on correct specification of model $\mathcal{B}$ (see e.g. Almirall, Ten Have and Murphy (2010), Henderson, Ansell and Alshibani (2010)) tend to give more efficient estimators (under correct model specification), the concern for misspecification of model $\mathcal{B}$ may be be considerable in view of the aforementioned difficulty of postulating this model. This distribution can indeed be difficult to specify in view of its multivariate nature, the fact that $U_m(\psi^*)$ represents a transformation of the observed data and that it may moreover share the same outcome over multiple time points, so that the models for $U_m(\psi^*)$ corresponding to different time points may not be congenial at all times. This concern can be overcome by inferring the conditional expectations $E\left\{U_m(\psi^*) \mid \overline{L}_m, \overline{A}_{m-1}\right\}$ from models for the conditional distribution of L_{m+1} given $\overline{L}_m, \overline{A}_{m-1}$ at each time m (Robins, Rotnitzky and Scharfstein, 2000; Almirall, Ten Have and Murphy, 2010). However, when the covariate $\overline{L}_m$ is high-dimensional and/or strongly associated with treatment A_m, specifying such models can be a thorny and non-trivial task.

6.2 Departures from sequential ignorability and sensitivity analysis

Specified departures from (32) can also yield identification. For instance, one can allow dependence of treatment on a specified portion of the future potential outcomes, by relaxing (32) to

$$A_m \perp\!\!\!\perp \underline{Y}_{m+\Delta}^{\overline{a}_{m-1},0} \mid \overline{L}_m, \overline{A}_{m-1} = \overline{a}_{m-1}, Y_{m+1:m+\Delta-1}^{\overline{a}_{m-1},0},$$

for some integer $\Delta \geq 1$; such assumptions have been termed future ignorability, since the independence at m is conditional on potential outcomes referring to times after m. This assumption can sometimes eliminate residual confounding bias, for instance, because the treatment process occurs in continuous time but confounding covariates are only measured intermittently, as is common in observational studies (Zhang, Joffe and Small, 2011), or when the future potential outcomes serve as proxies for other unmeasured confounding variables (Rosenbaum, 1984). However, it does not lead to nonparametric identification of the SNM parameters, so that inference becomes more dependent on correct specification of the causal model.

Alternatively, deviations from sequential ignorability can be parameterized as

$$\begin{aligned}(35)\qquad & f(A_m = a_m \mid \overline{L}_m = \overline{l}_m, \overline{A}_{m-1} = \overline{a}_{m-1}, \underline{Y}_{m+1}^{\overline{a}_{m-1},0} = \underline{y}_{m+1}) \\ &= \frac{t(A_m = a_m \mid \overline{L}_m = \overline{l}_m, \overline{A}_{m-1} = \overline{a}_{m-1}) \exp\left\{q_m(\underline{y}_{m+1}, \overline{l}_m, \overline{a}_m)\right\}}{\int t(A_m = a_m^\dagger \mid \overline{L}_m = \overline{l}_m, \overline{A}_{m-1} = \overline{a}_{m-1}) \exp\left\{q_m(\underline{y}_{m+1}, \overline{l}_m, (\overline{a}_{m-1}, a_m^\dagger))\right\} da_m^\dagger},\end{aligned}$$

with $q_m(\cdot)$ known, satisfying $q_m(\underline{y}_{m+1}, \overline{l}_m, \overline{a}_{m-1}, a_m = 0) = 0$ for all $(\underline{y}_{m+1}, \overline{l}_m, \overline{a}_{m-1})$ and with $t(A_m \mid \overline{L}_m, \overline{A}_{m-1})$ an unknown conditional density. With $q_m(\cdot) = 0$ encoding the assumption of sequential ignorability, the function $q_m(\cdot)$ thus expresses

the degree of departure from that assumption. As the data carry no genuine information about it, progress must be made by repeating the analysis with $q_m(\cdot)$ fixed at different values, which are then varied over some plausible range (Robins, Rotnitzky and Scharfstein, 2000); e.g. by setting $q_m(\underline{y}_{m+1}, \bar{l}_m, \bar{a}_m) = \gamma y_{m+1} a_m$, where γ is varied between -1 and 1.

6.3 Instrumental variables assumptions

When the assumption of sequential ignorability fails, progress can sometimes be also made using an instrumental variable (IV). Such variable A_0 is assumed to satisfy

$$A_0 \perp\!\!\!\perp \underline{Y}^0 \mid L_0 \tag{36}$$

and

$$F_{\underline{Y}^0 \mid L_0 = l_0, A_0 = a_0}\left(\underline{y}\right) = F_{\underline{Y}^{a_0,0} \mid L_0 = l_0, A_0 = a_0}\left(\underline{y}\right). \tag{37}$$

for all a_0, l_0 (Robins, 1989). Both these assumptions together imply that the instrument A_0 is not associated with the outcome, except through its association with subsequent treatments $A_m, m \geq 1$, which may affect outcome. These or similar assumptions have been used in adjusting for noncompliance in randomized trials (Robins and Tsiatis, 1991; Mark and Robins, 1993; Robins, 1994). With $A_m, m \geq 1$ denoting actual treatment and A_0 denoting randomised treatment, these assumptions are plausible when randomisation does not affect the outcome other than by influencing the actual treatment.

Estimation under the IV assumptions can be based on estimating equations (33) and (34), but requires setting $d_m(\overline{L}_{im}, \overline{A}_{im}) = 0$ and $d_m\left\{U_{im}(\psi), \overline{A}_{im}, \overline{L}_{im}\right\} = 0$ for $m > 0$. Because of these restrictions, root-n estimation of ψ^* typically requires additional assumptions on $\gamma_m^*(\bar{l}_m, \bar{a}_m; \psi^*)$ and $\gamma_m(\underline{y}_{m+1}, \bar{l}_m, \bar{a}_m; \psi^*)$. In particular, it is commonly assumed that these functions are linear in a_m and do not involve a_0; moreover, time-varying covariates are commonly ignored, i.e. L_m is set empty for $m > 0$. For instance, in linear SMMs for a single treatment A_1 (i.e., when $K = 1$) and dichotomous instrument, $\omega(L_0)$ in

$$E(Y_1 - Y_1^{a_0 0} \mid L_0 = l_0, \overline{A}_1 = \bar{a}_1) \quad = \quad \omega(l_0) a_1,$$

is just identified. Thus residual dependencies on a_0 or nonlinear dependencies on a_1 cannot be identified unless other untestable assumptions are imposed.

The resulting class of G-estimators contains the popular two-stage least squares estimator as a special case. However, the framework of G-estimation for SNMMs and SNDMs has the advantage that it extends immediately to outcomes that do not lend themselves to linear modelling, e.g. censored failure-time outcomes (Robins and Tsiatis, 1991) and dichotomous outcomes (Vansteelandt and Goetghebeur, 2003; Robins and Rotnitzky, 2004), as well as to sequential treatments (Robins and Hernan, 2009). For instance, when $K = 1$ and L_1 is empty, the logistic SMM

$$\frac{\text{odds}(Y_2^{a_1} = 1 \mid L_0 = l_0, \overline{A}_1 = \bar{a}_1)}{\text{odds}(Y_2^0 = 1 \mid L_0 = l_0, \overline{A}_1 = \bar{a}_1)} = \exp\left(\psi^* a_1\right),$$

can be fitted by solving the SMM estimating equations with $U^*(\psi)$ given by $\text{expit}\left\{\text{logit}E\left(Y_2 \mid \overline{A}_1, L_0\right) - \psi A_1\right\}$ (cfr. (3)) and $E\left(Y_1 \mid \overline{A}_1, L_0\right)$ substituted by the fitted value under a parametric model (Vansteelandt and Goetghebeur, 2003; Vansteelandt et al., 2011). This additional model may sometimes not be congenial with the SMM and instrumental variables assumptions in the sense that there may be no choice of parameter values indexing this model that satisfies these assumptions. This can be overcome by avoiding parameterization of the main effect of A_0 (conditional on L_0) in the model for $E\left(Y_1 \mid \overline{A}_1, L_0\right)$ and instead modelling the distribution of A_1, given A_0 and L_0 (Robins and Rotnitzky, 2004), or by completely saturating the parameterization of the main effect of A_0 (conditional on L_0) (Vansteelandt et al., 2011). van der Laan, Hubbard and Jewell (2007) abandon logistic SMMs in favor of an interesting, but difficult to interpret relative risk parameterization. Alternatively, multiplicative SMMs can be used; under such models, case-only estimators have been constructed, which remain valid under case-control sampling (Bowden and Vansteelandt, 2011).

Variant assumptions have been proposed that allow use of time-varying instruments along with SNMS and G-estimation. Robins and Hernan (2009) consider settings in which, at each time point, there is a variable whose association with the outcome of interest may be explained solely by its association with prior history and its effect on some treatment of interest. Joffe et al. (2010) consider settings in which the conditional independence of treatment and future potential outcomes in (28) holds for only an identifiable subset $\{i, m\}$ of the person-observations in the population rather than for all such observations. Treatment assignment in that subset may thus be considered an instrument for its effect and the effect of subsequent treatments.

IV analyses have several drawbacks relative to those based on sequential ignorability: 1) nonparametric identification is lost, and so inference is more dependent on correct specification of the causal model; 2) decreased power and precision; and (3) larger finite-sample bias.

6.4 Censoring

In SNFTMs, Type I censoring can be dealt with as previously explained by substituting $U_m(\psi)$ by an arbitrary function of $X_m(\psi) \equiv \min\{U_m(\psi), C_m(\psi)\}$ and $\Delta_m(\psi) \equiv I\{U_m(\psi) < C_m(\psi)\}$, where

$$C_m(\psi) \equiv \min\left\{U_m(C, \overline{a}_C, \overline{l}_C; \psi); \overline{a}_C, \overline{l}_C \in LA_m(C)\right\},$$

where $LA_m(C)$ is a given set of $(\overline{a}_C, \overline{l}_C)$ histories which agree with the observed history of L through time t_m or C, whichever comes first, and A through time t_{m-1} or C, whichever comes first, and where $U_m(C, \overline{a}_C, \overline{l}_C; \psi)$ is defined like $U_m(\psi)$ in Section 5.2, but with C replacing T and $\overline{a}_C$ and $\overline{l}_C$ replacing $\overline{A}_T$ and $\overline{L}_T$.

7. PREDICTING THE EFFECTS OF INTERVENTIONS

Identities (24) and (30) suggest using $U_m^*(\psi^*)$ and $U_m(\psi^*)$, respectively, as a prediction of $\underline{Y}_{m+1}^{\overline{a}_{m-1},0}$ among individuals with observed history $\overline{A}_m = \overline{a}_m$. In particular, $E\left(\underline{Y}^0\right) = E\{U_0^*(\psi^*)\}$ in SNMMs and $E\left(\underline{Y}^0\right) = E\{U_0(\psi^*)\}$ in SNDMs so that the expected outcome in the absence of treatment can be estimated as

the sample average of $U_0^*(\hat{\psi})$ in SNMMs and of $U_0(\hat{\psi})$ in SNDMs. To estimate $E\left(\underline{Y}^{\overline{a}_K}\right)$ for a different treatment regime $\overline{a}_K$, one could use a different structural nested model (SNM) with $\overline{a}_K$ as the reference treatment regime. However, when - as often - the interest lies in comparing the expected counterfactual outcomes between different treatment regimes, then a concern is that these different SNMs may fail to imply a coherent model. Further complications arise when the goal is to evaluate the expected counterfactual outcome following a dynamic treatment regime whereby the treatment at each time t_m is assigned as a function of the treatment and covariate history up to that time; that is, for each m, $a_m = g(\overline{a}_{m-1}, \overline{l}_m)$.

These complications can be overcome by supplementing the SNM with so-called current treatment interaction functions (Robins, Rotnitzky and Scharfstein, 2000) about which the data carry no information, but which enable one to transport treatment effects in the treated to population-averaged treatment effects. For instance, let $K = 1$ and suppose that a SNMM has been fitted with $g(\cdot)$ the identity link. For simplicity, we focus here only on the effect of a non-dynamic regime (a_0, a_1) at an end-of-study outcome $Y = Y_2$; results for dynamic treatment regimes are recovered upon making the substitutions $(g_0(l_0), g_1(a_0, \overline{l}_1))$ for (a_0, a_1). Two current treatment interaction functions can be defined, one for each sequential treatment:

$$\begin{aligned} r_1^*\left(\overline{L}_1, \overline{a}_1\right) &= E\left(Y^{a_0 a_1} - Y^{a_0 0} \mid \overline{A}_1 = \overline{a}_1, \overline{L}_1\right) \\ &\quad - E\left(Y^{a_0 a_1} - Y^{a_0 0} \mid A_0 = a_0, A_1 \neq a_1, \overline{L}_1\right) \\ r_0^*\left(L_0, \overline{a}_1\right) &= E\left(Y^{a_0 a_1} - Y^0 \mid A_0 = a_0, L_0\right) - E\left(Y^{a_0 a_1} - Y^0 \mid A_0 \neq a_0, L_0\right) \end{aligned}$$

These express how much the effects of subsequent treatment at m (i.e., a_1 and (a_0, a_1) at times 1 and 0, respectively) differ between groups that received that level of treatment at m and those that did not. Under the SNMM, it is easily deduced from knowledge of $r_1^*\left(\overline{L}_1, \overline{a}_1\right)$ and $r_0^*\left(L_0, \overline{a}_1\right)$ that $E\left(Y^{a_0 a_1} - Y^0 \mid A_0 = a_0, L_0\right)$ equals

$$\begin{aligned} & E\left(Y^{a_0 a_1} - Y^{a_0 0} \mid A_0 = a_0, L_0\right) + E\left(Y^{a_0 0} - Y^0 \mid A_0 = a_0, L_0\right) \\ &= E\left\{\gamma_1^*(\overline{L}_1, \overline{a}_1; \psi^*) - r_1^*\left(\overline{L}_1, \overline{a}_1\right) P(A_1 \neq a_1 \mid A_0 = a_0, \overline{L}_1) \mid A_0 = a_0, L_0\right\} \\ &\qquad + \gamma_0^*(L_0, a_0; \psi^*). \end{aligned}$$

Because $E\left(Y^{a_0 a_1} - Y^0 \mid L_0\right)$ moreover equals

$$E\left(Y^{a_0 a_1} - Y^0 \mid A_0 = a_0, L_0\right) - r_0^*\left(L_0, \overline{a}_1\right) P(A_0 \neq a_0 \mid L_0),$$

we thus obtain that $E\left(Y^{a_0 a_1}\right) = E\left(Y^{a_0 a_1} - Y^0\right) + E\left(Y^0\right)$ equals

$$\begin{aligned} & E\left[E\left\{\gamma_1^*(\overline{L}_1, \overline{a}_1; \psi^*) - r_1^*\left(\overline{L}_1, \overline{a}_1\right) P(A_1 \neq a_1 \mid A_0 = a_0, \overline{L}_1) \mid A_0 = a_0, L_0\right\}\right. \\ & \left. + \gamma_0^*(L_0, a_0; \psi^*) - r_0^*\left(L_0, \overline{a}_1\right) P(A_0 \neq a_0 \mid L_0) + U_0(\psi^*)\right]. \end{aligned}$$

When there is no current treatment interaction (i.e. $r_1^*\left(\overline{l}_1, \overline{a}_1\right) = r_0^*\left(l_0, \overline{a}_1\right) = 0$ for all a_0, a_1, l_0, l_1), we thus have that

$$E\left(Y^{a_0 a_1}\right) = E\left[E\left\{\gamma_1^*(\overline{L}_1, \overline{a}_1; \psi^*) \mid A_0 = a_0, L_0\right\} + \gamma_0^*(L_0, a_0; \psi^*) + U_0(\psi^*)\right].$$

While the components $\gamma_1^*(\overline{L}_1, \overline{a}_1; \psi^*)$, $\gamma_0^*(L_0, a_0; \psi^*)$ and $U_0(\psi^*)$ can be estimated along the lines described in previous sections, a complication is that a model for the distribution of L_1, conditional on A_0, L_0, is needed to evaluate this; this can be cumbersome when L_1 is high-dimensional. In simple structural models in which there is no effect modification by post-treatment variables (i.e., $\gamma_1^*(\overline{L}_1, \overline{a}_1)$ is not a function of L_1) and non-dynamic regimes are considered, this complication is avoided.

The assumption of no current treatment interaction is satisfied under a mild strengthening of sequential ignorability such that:

$$A_m \perp\!\!\!\perp \underline{Y}_{m+1}^{\overline{a}_K} \mid \overline{L}_m, \overline{A}_{m-1} = \overline{a}_{m-1},$$

for all m and all treatment histories $\overline{a}_K$. It is likewise sometimes satisfied under a mild strengthening of the instrumental variables assumption (36) such that for all treatment histories $\overline{a}_K$:

$$A_0 \perp\!\!\!\perp \underline{Y}^{\overline{a}_K} \mid L_0.$$

Indeed, suppose for instance that A_1 is binary (0/1) and that

$$E\left(Y^{a_0 a_1} - Y^{a_0 a_1^\dagger} \mid A_1 = a_1, A_0 = a_0, L_0\right) = \gamma_1^*(a_1^\dagger, L_0; \psi^*)(a_1 - a_1^\dagger),$$

for all $a_1, a_1^\dagger$. The following the instrumental variables assumptions, $Y^{a_0 0}$ and $Y^{a_0 1}$ should then be independent of A_0, given L_0, which respectively implies that

$$\begin{aligned} E\left\{Y - \gamma_1^*(0, L_0; \psi^*)A_1 \mid A_0, L_0\right\} &= E\left\{Y - \gamma_1^*(0, L_0; \psi^*)A_1 \mid L_0\right\} \\ E\left\{Y - \gamma_1^*(1, L_0; \psi^*)(1 - A_1) \mid A_0, L_0\right\} &= E\left\{Y - \gamma_1^*(1, L_0; \psi^*)(1 - A_1) \mid L_0\right\}. \end{aligned}$$

It follows from this that $\gamma_1^*(0, L_0; \psi^*) = -\gamma_1^*(1, L_0; \psi^*)$ and thus again that the no current treatment interaction assumption is satisfied (Hernan and Robins, 2006).

8. DIRECT AND INDIRECT EFFECTS

SNMs parameterize the effects of treatment at each time with subsequent treatments set to some reference level. These effects can be viewed as controlled direct effects (Robins and Greenland, 1992), controlling all subsequent treatments at their reference levels. The formalism of SNMs is therefore more widely applicable for inferring the controlled direct effect of some target exposure A_0 on an outcome Y, other than through some mediator A_1 (e.g. the direct effect of the FTO gene on the risk of myocardial infarction other than via body mass). In particular, in the SNMM

$$\begin{aligned} E\left(Y - Y^{a_0 0} \mid \overline{A}_1 = \overline{a}_1, \overline{L}_1\right) &= \gamma_1^*\left(\overline{a}_1, \overline{L}_1; \psi^*\right) \\ E\left(Y^{a_0 0} - Y^0 \mid A_0 = a_0, L_0\right) &= \gamma_0^*\left(a_0, L_0; \psi^*\right), \end{aligned}$$

$\gamma_0^*\left(a_0, L_0; \psi^*\right)$ encodes the controlled direct effect of setting A_0 to zero, controlling A_1 at zero uniformly in the population. However, caution is warranted because $\gamma_0^*\left(a_0, L_0; \psi^*\right)$ may not encode the controlled direct effect of setting A_0 to zero, when controlling A_1 at some value $a_1 \neq 0$ (Robins and Wasserman, 1997). Robins (1999) thus proposed directly parameterizing the controlled direct effect as

$$E\left(Y^{a_0 a_1} - Y^{0 a_1} \mid A_0 = a_0, L_0\right) = m(a_0, a_1, L_0; \psi^*), \tag{38}$$

where $m(a_0, a_1, L_0; \psi)$ is a known function, smooth in ψ, which satisfies $m(0, a_1, l_0; \psi) = 0$. In contrast to SNMMs, (38) parameterizes only the effect of a_0; in (38), a_1 may, however, be a modifier of the effect of a_0.

Since model (38) for fixed a_1 is a SMM for the counterfactual outcome Y^{a_1}, the techniques of Section 3 would be applicable to estimate ψ^* if Y^{a_1} were observed for each subject. Since Y^{a_1} is only observed for individuals with exposure level a_1, Robins (1999) proposed treating subjects who receive a level of A_1 other than a_1 as censored and, assuming sequential ignorability, to inversely weight the data by the density $f(A_1 \mid A_0, \overline{L}_1)$ to control resulting selection bias. This amounts to solving ψ from an estimating equation of the form

$$(39) \qquad 0 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{f(A_{i1} \mid A_{i0}, \overline{L}_{i1})} \left[d(A_{i0}, L_{i0}) - E\left\{d(A_{i0}, L_{i0}) \mid L_{i0}\right\}\right] \times \left[Y_i - m(\overline{A}_{i1}, L_{i0}; \psi) - E\left\{Y_i - m(\overline{A}_{i1}, L_{i0}; \psi) \mid L_{i0}\right\}\right],$$

where $d(A_{i0}, L_{i0})$ is an arbitrary index function. More efficient and doubly robust estimators have been reported elsewhere (Goetgeluk, Vansteelandt and Goetghebeur, 2008), as well as extensions to time-varying treatments (Robins, 1999).

Ignorability assumptions can be most compelling in randomized trials (and Mendelian randomization studies), where assumption (36) is guaranteed by design, whereas the processes underlying the evolution of subsequent mediators may be poorly understood. Robins and Greenland (1994) used initial randomization (or more generally, instrumental variables assumptions) to estimate controlled direct effects with SNFTMs. One can also use these approaches with SNMMs or SNDMs (e.g., Ten Have et al., 2007) and, in principle, in the presence of multiple mediators.

SMMs have also been developed for so-called natural direct effects (Robins and Greenland, 1992; Pearl, 2001). With $Y^{a_0 A_1^0}$ denoting the counterfactual outcome if A_0 were set to a_0 and A_1 to the counterfactual level A_1^0 that A_1 would take if A_0 were set to zero, these are defined by contrasts between $Y^{a_0 A_1^0}$ and $Y^{0 A_1^0}$ for some $a_0 \neq 0$. Because A_1^0 may often reflect a natural level of A_1 (as in the absence of treatment) which differs between subjects, natural direct effects may have a more appealing interpretation than controlled direct effects. They moreover correspond with a measure of natural indirect effect in terms of contrasts between $Y^{a_0 A_1^{a_0}}$ and $Y^{a_0 A_1^0}$ for some $a_0 \neq 0$. SMMs for natural direct effects have been considered van der Laan and Petersen (2008) and Tchetgen Tchetgen and Shpitser (2011). Such models are defined by

$$(40) \qquad E\left(Y^{a_0 A_1^0} - Y^{0 A_1^0} \mid A_0 = a_0, L_0 = l_0\right) = m(a_0, l_0; \psi^*),$$

for each a_0, l_0, where $m(a_0, L_0; \psi)$ is a known function, smooth in ψ, which satisfies $m(0, L_0; \psi) = 0$. Extensions to sequential treatments or mediators have so far not been developed in view of difficulties of identification in such settings.

9. CONCLUDING REMARKS

Structural nested models were designed in part to deal with confounding by variables affected by treatment. These models maintain close resemblance to ordinary regression models by parameterizing conditional treatment effects. However, in contrast to ordinary regression models, they avoid conditioning on post-treatment variables by modelling the outcome at each time conditional on the

treatment and covariate history up to that time; they do this after having removed the effects of later treatments so as to disentangle the unique contributions of each treatment at each time. The associated method of G-estimation has close resemblance to ordinary regression methods because it realizes control for measured confounders through conditioning. In spite of these strong connections, SNMs and G-estimation have not become quite as popular as MSMs and the associated IPW methods (Robins, Hernan and Brumback, 2000).

The lack of popularity of G-estimation is largely related to the fact that it cannot usually be performed via off-the-shelf software; however, note that SAS and Stata macros for SNFTMs and SNCFTMs are available at `http://www.hsph.harvard.edu/causal/software/`. This lack of popularity is additionally related to difficulties in solving the estimating equations in the analysis of censored survival times using SNFTMs. These difficulties can now be overcome by using the newer class of SNCFTMs instead (Picciotto et al., 2011; Martinussen et al., 2011). In spite of these limitations, SNMs and G-estimation allow for greater flexibility than MSMs and typically yield better performing estimators (see Section 4.1). This is especially so when handling continuous exposures or when handling a binary exposure that is strongly correlated with subject characteristics (e.g., when the treated and untreated are very different in terms of subject characteristics). In the latter case, IPW estimators will typically have a poor performance, reflecting the lack of information about the treatment effect in strata where most/all subjects are treated or untreated. In contrast, because SNMs parameterize treatment effects conditionally on covariates, non-saturated models allow for borrowing of information, so that G-estimators can pool the treatment effects across strata, as in expression (23), downweighing those strata where information on treatment effect is lacking. SNMs can also incorporate effect modification by time-varying covariates. As such, a saturated SNM encodes all possible causal contrasts on the considered scale, in contrast to MSMs which average the effects across (time-varying) covariates, thereby diluting the effects when effect heterogeneity exists on the considered scale. SNMs can moreover make use of instrumental variables.

G-estimation is not to be confused with G-computation (Robins, 1986), which involves standardizing the predictions from an outcome regression model corresponding to the considered treatment regime, relative to the confounder distribution in the population. Up to recently, also this approach has received little attention in practice because it is computationally intensive and because correct specification of models for the distribution of the (possibly high-dimensional) confounders can be a thorny issue in practice. These concerns, which also relate to likelihood-based inference under SNDMs (Robins, Rotnitzky and Scharfstein, 2000), can be somewhat mitigated by summarising the confounders at each time by a longitudinal propensity score defined as the probability of treatment at that time, given the history of confounders at that time (Achy-Brou, Frangakis and Griswold, 2010). However, this may demand correct specification of propensity score models in addition to a model for the outcome at each time. G-computation moreover does not enable a transparent parameterization of the effect of a particular treatment regime on the outcome and may thereby imply a null paradox (Robins and Wasserman, 1999) according to which tests of the null hypothesis of no effect may be guaranteed to reject in large samples (Robins, 1997). However, recent empirical applications have turned out to be rather successful (see e.g.

Cain et al., 2011).

We have attempted to make the literature on structural nested models and G-estimation more accessible, while also giving pointers to the related literatures on effect modification and mediation. Variants of SNMs have also been developed to help identify optimal sequences of treatments when treatments may be assigned dynamically as a function of previous treatment and covariate history. In such settings, it is more natural to model the effect of a blip of treatment at m on a particular utility function Y, such as the outcome at the end-of-study time, if all subsequent treatments are optimal; i.e. $a_k = a_k^{\text{opt}}(\bar{l}_k, \overline{a}_{k-1})$ for $k > m$. This can be done by parameterizing the so-called regrets: contrasts of $E\left(Y^{\overline{a}_m, \underline{a}_{m+1}^{\text{opt}}} \mid \overline{L}_m, \overline{A}_m = \overline{a}_m\right)$ and $E\left(Y^{\overline{a}_{m-1}, \underline{a}_m^{\text{opt}}} \mid \overline{L}_m, \overline{A}_m = \overline{a}_m\right)$ (Murphy, 2003). Alternatively, since the optimal treatment is unknown, it may be easier to parameterize the effect of a blip of treatment at m *relative to no treatment* when all future treatments are optimal. This amounts to contrasting $E\left(Y^{\overline{a}_m, \underline{a}_{m+1}^{\text{opt}}} \mid \overline{L}_m, \overline{A}_m = \overline{a}_m\right)$ and $E\left(Y^{\overline{a}_{m-1}, 0, \underline{a}_{m+1}^{\text{opt}}} \mid \overline{L}_m, \overline{A}_m = \overline{a}_m\right)$ (Robins, 2004). We refer the reader to other papers in this issue for detailed accounts of such models. We conclude by expressing our hope that continued efforts to develop computational algorithms and corresponding software programs for SNMs will aid to make these methods accessible to a wider audience.

ACKNOWLEDGEMENTS

The authors are grateful to the editors and reviewers for very detailed feedback which substantially improved an earlier version of this manuscript. The first author acknowledges support from the Flemish Research Council (FWO) research grant G.0111.12 and IAP research network grant nr. P07/05 from the Belgian government (Belgian Science Policy). The second author acknowledges support from the US NIH (grants # R01-DK090385, RC4-MH092722, and R01-MH078016).

REFERENCES

Achy-Brou, A. C., Frangakis, C. E., and Griswold, M. (2010). Estimating treatment effect of longitudinal designs using regressin models on propensity scores. *Biometrics* **66**, 824-833.

Almirall, D., Ten Have, T. and Murphy, S.A. (2010). Structural Nested Mean Models for Assessing Time-Varying Effect Moderation. *Biometrics* **66**, 131-139.

Bowden, J. and Vansteelandt, S. (2011). Mendelian randomisation analysis of case-control data using Structural Mean Models. *Statistics in Medicine* **30**, 678-694.

Cain, L.E., Logan, R., Robins, J.M., Sterne, J.A.C., Sabin, C., Bansi, L., Justice, A., Goulet, J., van Sighem, A., de Wolf, F., Bucher, H.C., von Wyl, V., Esteve, A., Casabona, J., del Amo, J., Moreno, S., Seng, R., Meyer, L., Perez-Hoyos, S., Muga, R., Lodi, S., Lanoy, E., Costagliola, D., Hernan, M.A. (2011). When to Initiate Combined Antiretroviral Therapy to Reduce Mortality and AIDS-Defining Illness in HIV-Infected Persons in Developed Countries: An Observational Study. *Annals of Internal Medicine* **154**, 509-W173.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* **34**, 305-334.

Frangakis, C.A. and Rubin, D.B. (2002). Principal Stratification in Causal Inference. *Biometrics*, **58**, 21-29.

Greenland, S. and Robins, J.M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology* **15**, 412-418.

Greenland, S., Robins, J. and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science* **14**, 29-46.

Goetgeluk, S., Vansteelandt, S. and Goetghebeur, E. (2008). Estimation of controlled direct effects. *Journal of the Royal Statistical Society Series B* **70**, 1049-1066.

Henderson, R., Ansell, P. and Alshibani, D. (2010). Regret-Regression for Optimal Dynamic Treatment Regimes. *Biometrics* **66**, 1192-1201.

Hernan, M.A. and Robins, J.M. (2006). Instruments for causal inference - An epidemiologist's dream? *Epidemiology* **17**, 360-372.

Hernan, M.A. (2010). The hazards of hazard ratios. *Epidemiology* **21**, 13-15.

Joffe, M.M., Small, D. and Hsu, C.-Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science* **22**, 74-97.

Joffe, M.M., Yang, W.P. and Feldman, H. (2012), G-Estimation and Artificial Censoring: Problems, Challenges, and Applications. *Biometrics* **68**, 275-286.

Mark, S.D. and Robins, J.M. (1993). A method for the analysis of randomized trials with compliance information: an application to the multiple risk factor intervention trial. *Controlled Clinical Trials* **14**, 79-97.

Martinussen, T., Vansteelandt, S., Gerster, M. and Hjelmborg, J.v.B. (2011). Estimation of direct effects for survival data using the Aalen additive hazards model. *Journal of the Royal Statistical Society - Series B* **73**, 773-788.

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

Newey, W.K. (1990). Semiparametric Efficiency Bounds. *Journal of Applied Econometrics* **5**, 99-135.

Pearl, J. (1995). Causal Diagrams for Empirical Research (with discussion). *Biometrika* **82**, 669-710.

Pearl, J. (2001). Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, CA, 411-420.

Picciotto, S., Hernan, M.A., Page, J., Young, J.G. and Robins, J.M. (2012). Structural Nested Cumulative Failure Time Models to Estimate the Effects of Interventions. *Journal of the American Statistical Association*, **107**, 886-900.

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

Robins, J.M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health Service Research Methodology: A Focus on AIDS.* Eds: Sechrest L., Freeman H., Mulley A. Washington, D.C.: U.S. Public Health Service, National Center for Health Services Research., pp. 113-159.

Robins, J.M., and Tsiatis, A.A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications in Statistics, Theory and Methods* **20**, 2609-2631.

Robins, J.M. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* **79**, 321-334.

Robins, J.M., Blevins, D., Ritter, G. and Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystic carinii pneumonia on the survival of AIDS patients. *Epidemiology* **3**, 319-336.

Robins, J.M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143-155.

Robins, J.M., Mark, S.D. and Newey, W.K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48**, 479-495.

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

Robins, J.M., and Greenland, S. (1994). Adjusting for differential rates of prophylaxis therapy for PCP in high- versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association* **89**, 737-749.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression-coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846-866.

Robins, J.M. (1997). *Causal Inference from Complex Longitudinal Data.* Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics (120), M. Berkane, Editor. NY: Springer Verlag, pp. 69-117.

Robins, J.M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, **16**, 285-319.

Robins, J.M. and Wasserman, L. (1997). *Estimation of Effects of Sequential Treatments by Reparameterizing Directed Acyclic Graphs.* Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, Providence Rhode Island, August 1-3, 1997. Dan Geiger and Prakash Shenoy (Eds.), Morgan Kaufmann, San Francisco, pp. 409-420.

Robins, J.M. (1999). *Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models.* In: Computation, Causation, and Discovery. Eds. C. Glymour and G. Cooper. Menlo Park, CA, Cambridge, MA: AAAI Press/The MIT Press, pp. 349-405.

Robins, J.M., Rotnitzky, A., and Scharfstein, D.O. (2000). *Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models,* in Statistical Models in Epidemiology, eds. E. Halloran and D. Berry, New York: Springer-Verlag.

Robins, J.M. (2000). *Marginal structural models versus structural nested models as tools for causal inference.* In Statistical Models in Epidemiology: The Environment and Clinical Trials, Ed. M. Halloran and D. Berry, New York: Springer-Verlag, pp. 95134.

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

Robins, J.M. and Rotnitzky, A. (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. (2004). *Optimal structural nested models for optimal sequential decisions.* In DY Lin and P Heagerty (Eds.), Proceedings of the Second Seattle Symposium on Biostatistics, New York. Springer.

Robins, J.M. and Rotnitzky, A. (2004). Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika* **91**, 763-783.

Robins, J.M. and Hernan, M.A. (2009). *Estimation of the causal effects of time-varying exposures.* In: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G, eds. Longitudinal Data Analysis. Boca Raton, FL: CRC Press; pp. 553-599.

Rosenbaum, P.R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, A* **147**, 656-666.

Rosenbaum, P.R. and Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, **79**, 516-524.

Rubin, D.B. (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* **6**, 3458.

Stephens, A., Keele, L., Joffe, M. (2013). Estimating Post-Treatment Effect Modification With Generalized Structural Mean Models. Submitted.

Tchetgen Tchetgen, E., Robins, J.M. and Rotnitzky, A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika* **97**, 171-180.

Tchetgen Tchetgen, E. and Robins, J.M. (2010). The Semiparametric Case-Only Estimator. *Biometrics* **66**, 1138-1144.

Tchetgen Tchetgen, E.J. and Rotnitzky, A. (2011). Double-robust estimation of an exposure-outcome odds ratio adjusting for confounding in cohort and case-control studies. *Statistics in Medicine* **30**, 335347.

Tchetgen Tchetgen, E.J. and Shpitser, I. (2011). Semiparametric Estimation of Models for Natural Direct and Indirect Effects. Harvard University Biostatistics Working Paper Series. Working Paper 129.

Tchetgen Tchetgen, E.J. (2012). Multiple-Robust Estimation of an Odds Ratio Interaction. *Harvard University Biostatistics Working Paper Series. Working Paper 142.* http://biostats.bepress.com/harvardbiostat/paper142.

Ten Have, T.R., Joffe, M.M., Lynch, K.G., Brown, G.K., Maisto, S.A. and Beck, A.T. (2007). Causal mediation analyses with rank preserving models. *Biometrics* **63**, 926-934.

van der Laan, M.J., Hubbard, A. and Jewell, N. (2007). Estimation of treatment effects in randomized trials with non-compliance and a dichotomous outcome. *Journal of the Royal Statistical Society - Series B* **69**, 463-482.

van der Laan, M.J. and Petersen, M.L. (2008). Direct Effect Models. *The International Journal of Biostatistics* **4**, 1-27.

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

Vansteelandt, S., VanderWeele, T., Tchetgen, E.J. and Robins, J.M. (2008a). Semiparametric inference for statistical interactions. *Journal of the American Statistical Association* **103**, 16931704.

Vansteelandt, S., DeMeo, D., Su, J., Smoller, J., Murphy, A.J., McQueen, M., Celedon, J., Weiss, S.T., Silverman, E.K. and Lange, C. (2008b). Testing and estimating gene-environment interactions in family-based association studies. *Biometrics* **64**, 458-467.

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

Vansteelandt, S., Bowden, J., Babanezhad, M., and Goetghebeur, E. (2011). On instrumental variables estimation of causal odds ratios. *Statistical Science*, **26**, 403-422.

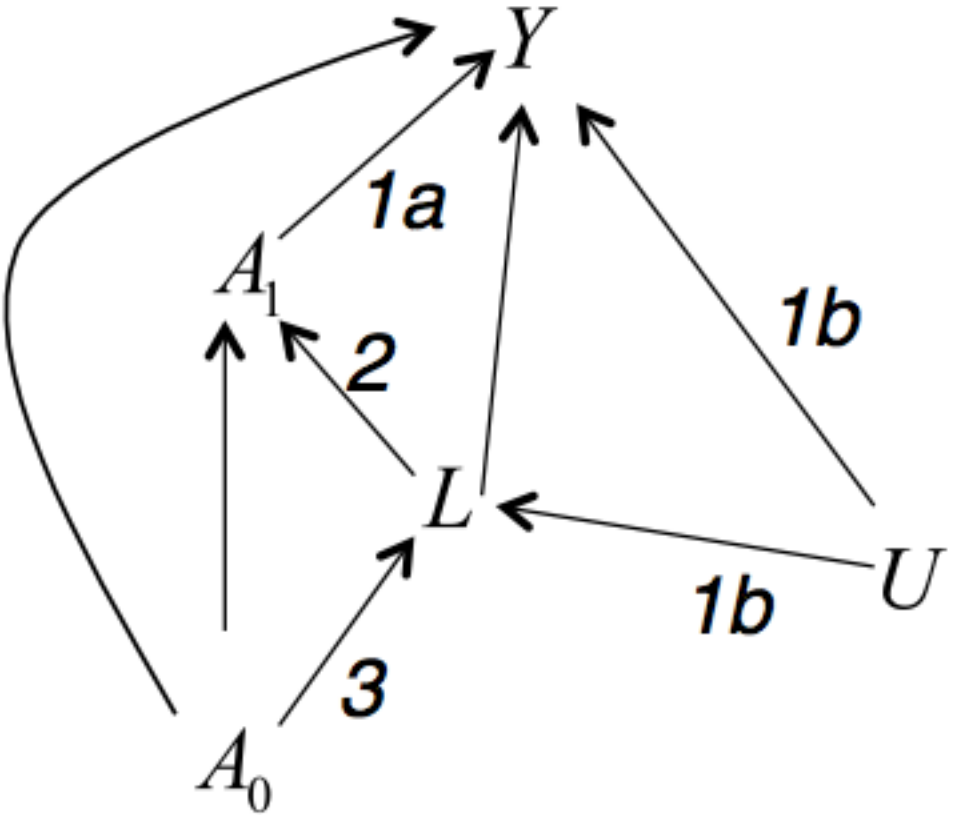
Vansteelandt, S., Bekaert, M. and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research* **21**, 7-30.

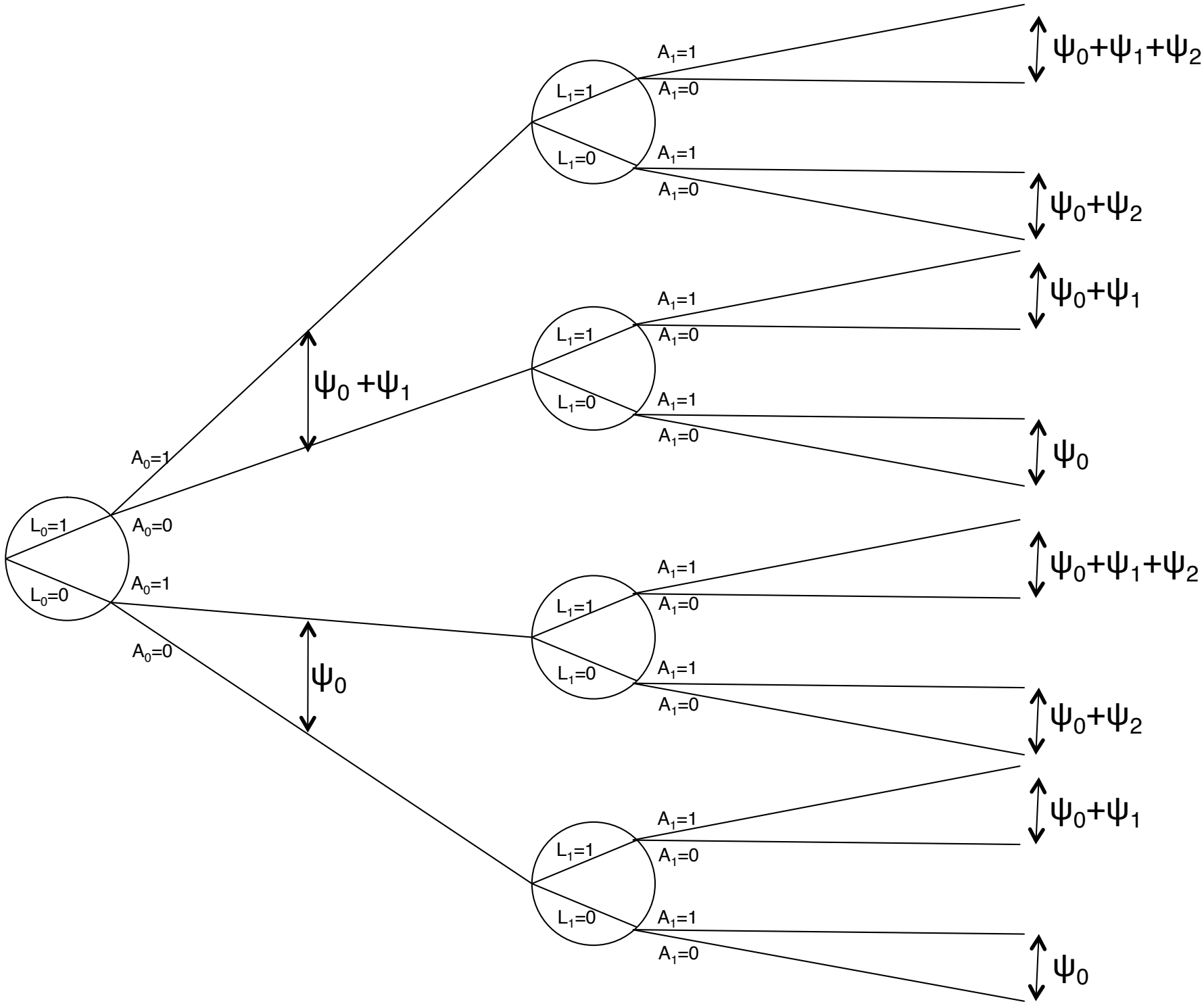
Vansteelandt, S. and Daniel, R.M. (2013). On regression adjustment for the propensity score. Submitted.

Vock, D.M., Tsiatis, A.A., Davidian, M., Laber, E.B., Tsuang, W.M., Finlen-Copeland, C.A. and Palmer, S.M. (2013). Assessing the causal effect of organ transplantation on the distribution of residual lifetime. *Biometrics*, in press.

Wei, L.J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine* **11**: 1871-1879.

Zhang, M., Joffe, M.M., and Small, D.S. (2011). Causal inference for continuous-time processes when covariates are observed only at discrete times. *Annals of Statistics* **39**, 131-173.


Y
1a
A_1
2
1b
L
U
1b
3
A_0


$L_0=1$
$L_0=0$
$A_0=1$
$A_0=0$
$A_0=1$
$A_0=0$
$\psi_0+\psi_1$
ψ_0
$L_1=1$
$L_1=0$
$A_1=1$
$A_1=0$
$\psi_0+\psi_1+\psi_2$
$\psi_0+\psi_2$
$\psi_0+\psi_1$
ψ_0
$\psi_0+\psi_1+\psi_2$
$\psi_0+\psi_2$
$\psi_0+\psi_1$
ψ_0