Reprinted with permission from
Latent Variable Modeling and Applications to Causality.
Lecture Notes in Statistics (120), M. Berkane, Editor.
NY: Springer Verlag, 69-117. 1997.
All rights reserved.

Causal Inference from Complex Longitudinal Data

James M. Robins
Departments of Epidemiology and Biostatistics
Harvard School of Public Health
Boston, MA 02115
email: robins@hsph.harvard.edu

1. Introduction

The subject-specific data from a longitudinal study consist of a string of numbers. These numbers represent a series of empirical measurements. Calculations are performed on these strings of numbers and causal inferences are drawn. For example, an investigator might conclude that the analysis provides strong evidence for "a direct effect of AZT on the survival of AIDS patients controlling for the intermediate variable – therapy with aerosolized pentamidine." The nature of the relationship between the sentence expressing these causal conclusions and the computer calculations performed on the strings of numbers has been obscure. Since the computer algorithms are well-defined mathematical objects, it is important to provide formal mathematical definitions for the English sentences expressing the investigator's causal inferences.

I proposed a formal theory of counterfactual causal inference (Robins, 1986, 1987) that extended Rubin's (1978) "point treatment" theory to longitudinal studies with direct and indirect effects and time-varying treatments, confounders, and concomitants. The purpose of this paper is to provide a summary and unification of this theory.

In my theory, causal questions that could be asked concerning the direct and indirect effects of the measured variables on an outcome became mathematical conjectures about the causal parameters of event trees that I called causally interpreted structured tree graphs (CISTGs). I defined randomized CISTGs (RCISTGs) and showed that for RCISTGs, a subset of the population causal parameters were non-parametrically identified by the G-computation algorithm formula. A RCISTG is an event tree in which the treatment received at time t is randomly allocated (ignorable) conditional on past treatment, outcome, and covariate history. In the absence of physical randomization, as in an observational study, the assumption that a CISTG is an RCISTG is non-identifiable.

Pearl (1995), and Spirtes, Glymour, and Scheines (SGS) (1993) recently developed a formal theory of causal inference based on causal directed acyclic graphs (DAGs). I show that these causal DAGs are the same mathematical objects as particular RCISTGs and thus a theorem in one causal theory is a theorem in the other (Robins, 1995b).

The standard approach to the estimation of a time-varying treatment on an outcome of interest is to model the probability of the outcome at time t as a function of past treatment history. This approach may be biased, whether or not one further adjusts for the past history of time dependent confounding covariates, when these covariates predict subsequent outcome and treatment history and are themselves influenced by past treatment. In this setting, I have proposed several methods that can provide unbiased estimates of the causal effect of a time-varying treatment in the presence of time varying confounding factors. In this paper, I describe two of these methods of estimation: estimation of the conditional probabilities in the G-computation algorithm formula (Robins 1986, 1987, 1989), and G-estimation of structural nested models (Robins 1989, 1992, 1993, 1994, 1995a, 1996). The G-computation algorithm formula is equivalent to the marginal distribution of the outcome in the manipulated subgraph of a DAG in which all arrows into the treatment variables are removed and the vector of treatment random variables X is set to a treatment regime or plan x of interest. This marginal distribution has a causal interpretation as the effect of treatment regime x on the outcome if the graph is a RCISTG, i.e., treatment received at each time t was randomly allocated (i.e., ignorable) conditional on past treatment and covariate history.

However, estimation of causal effects based on the G-computation algorithm is seriously non-robust since the inevitable misspecification of statistical models for the conditional law of the DAG variables given their parents results in bias under the causal null hypothesis of no treatment effect. In contrast, an approach based on estimation of structural nested models can often avoid this bias. Mathematically, structural nested models reparameterize the joint law of the DAG in terms of parameters that represent contrasts between the marginal distributions of the outcome in the manipulated graph with treatment X set to different values of x. Causally the parameters of a structural nested model represent the causal effect of a final brief blip of treatment on the outcome of interest. However, in Sec. 8.2, we show that structural nested models, like models for the conditional laws of the DAG variables given their parents, are non-robust for testing the null hypothesis of no direct effect of a time-dependent treatment X_1 controlling for a time-dependent intermediate variable X_2. Therefore, in Appendix 3, we introduce an extension of structural nested models, the “direct effect” structural nested models, that are suitable for testing the null hypothesis of no direct effect.

In Section 9, we discuss how one can use G-estimation to estimate the parameters of a structural nested model even when treatment at time t is not allocated at random given the past (i.e., the CISTG is not a RCISTG), provided that one can correctly specify a parametric or semiparametric model for the probability of treatment at t conditional on past covariate and treatment history and on the (possibly unobserved) value of a subject’s counterfactual outcome (Robins, 1996).

In Sections 1-9, we assume that treatment and covariate processes change (jump) at the pre-specified times. In Section 10 we relax this assumption by introducing continuous-time structural nested models whose parameter ψ reflects the instantaneous causal effect of the treatment rate (Robins, 1996). In Sec. 11, I argue that the faithfulness assumption of Pearl and Verma (1991) and SGS (1993) should not be used to draw causal conclusions from observational data.

2. Standard Analysis of Sequential Randomized Trials

The following type of example originally motivated the development of the methods described in this article. The graph in Figure 1 represents the data obtained from a hypothetical (oversimplified) sequential randomized trial of the joint effects of AZT (A_0) and aerosolized pentamidine (A_1) on the survival of AIDS patients. AZT inhibits the AIDS virus. Aerosolized pentamidine prevents pneumocystis pneumonia (PCP), a common opportunistic infection of AIDS patients. The trial was conducted as follows. Each of 32,000 subjects was randomized with probability .5 to AZT ($A_0 = 1$) or placebo ($A_0 = 0$) at time t_0. All subjects survived to time t_1. At time t_1, it was determined whether a subject had had an episode of PCP ($L_1 = 1$) or had been free of PCP ($L_1 = 0$) in the interval $(t_0, t_1]$. Since PCP is a potential life-threatening illness, all subjects with $L_1 = 1$ were treated with aerosolized pentamidine (AP) therapy ($A_1 = 1$) at time t_1. Among subjects who were free of PCP ($L_1 = 0$), one-half were randomized to receive

AP at t_1 and one half were randomized to placebo ($A_1 = 0$). At time t_2, the vital status was recorded for each subject with $Y = 1$ if alive and $Y = 0$ if deceased. We view A_0, L_1, A_1, Y as random variables with realizations a_0, l_1, a_1, y. All investigators agreed that the data supported a beneficial effect of treatment with AP ($A_1 = 1$) because, among the 8,000 subjects with $A_0 = 1$, $L_1 = 0$, AP was assigned at random and the survival rates were greater among those given AP, since

$$P\left[Y = 1 \mid A_1 = 1, L_1 = 0, A_0 = 1\right] - P\left[Y = 1 \mid A_1 = 0, L_1 = 0, A_0 = 1\right] = \\ 3/4 - 1/4 = 1/2 \tag{2.1}$$

The remaining question was whether, given that subjects were to be treated with AP, should or should not they also be treated with AZT. That is, we wish to determine whether the direct effect of AZT on survival controlling for (the potential intermediate variable) AP is beneficial or harmful (when all subjects receive AP). The most straightforward way to examine this question is to compare the survival rates in groups with a common AP treatment who differ on their AZT treatment. Reading from Figure 1 we observe, after collapsing over the data on L_1-status, that

$$P\left[Y = 1 \mid A_0 = 1, A_1 = 1\right] - P\left[Y = 1 \mid A_0 = 0, A_1 = 1\right] = \\ 7/12 - 10/16 = -1/24 \tag{2.2}$$

suggesting a harmful effect of AZT. However, the analysis in (2.2) fails to account for the possible confounding effects of the extraneous variable PCP(L_1). [We refer to PCP here as an “extraneous variable” because the causal question of interest, i.e., the question of whether AZT has a direct effect on survival controlling for AP, makes no reference to PCP. Thus adjustment for PCP is necessary only insofar as PCP is a confounding factor.] It is commonly accepted that PCP is a confounding factor and must be adjusted for in the analysis if PCP is *(a)* an independent risk (i.e., prognostic) factor for the outcome and *(b)* an independent risk factor for (predictor of) future treatment. By “independent” risk factor in *(a)* and *(b)* above, we mean a variable that is a predictor conditional upon all other measured variables occurring earlier than the event being predicted. Hence, to check condition *(a)*, we must adjust for A_0 and A_1; to check condition *(b)*, we must adjust for A_0.

Reading from Figure 1, we find that conditions *(a)* and *(b)* are both true, i.e.,

$$.5 = P\left[Y = 1 \mid L_1 = 1, A_0 = 1, A_1 = 1\right] \neq P\left[Y = 1 \mid L_1 = 0, A_0 = 1, A_1 = 1\right] = .75 \tag{2.3}$$

and

$$1 = P\left[A_1 = 1 \mid L_1 = 1, A_0 = 1\right] \neq P\left[A_1 = 1 \mid L_1 = 0, A_0 = 1\right] = .5 \tag{2.4}$$

The standard approach to the estimation of the direct effect of AZT controlling for AP in the presence of a confounding factor (PCP) is to compare survival rates among groups with common AP and confounder history (e.g., $L_1 = 1$, $A_1 = 1$) but who differ in AZT treatment. Reading from Figure 1, we obtain

$$P\left[Y = 1 \mid A_0 = 1, L_1 = 1, A_1 = 1\right] - P\left[Y = 1 \mid A_0 = 0, L_1 = 1, A_1 = 1\right] \\ = 4,000/8,000 - 10,000/16,000 = -1/8 \tag{2.5}$$

Hence the analysis adjusted for PCP also suggests an adverse direct effect of AZT on survival controlling for AP.

Figure 1 goes here.

Figure 1 is not presently in a computer-readable format. To obtain a hard copy, please contact Ms. Patricia Shechter (email: pats@epinet.harvard.edu).

However, the analysis adjusted for PCP is also problematic, since *(a)* Rosenbaum (1984) and Robins (1986, 1987) argue that it is inappropriate to adjust (by stratification) for an extraneous risk factor that is itself affected by treatment, and *(b)*, reading from Figure 1, we observe that PCP is affected by previous treatment, i.e.,

$$.5 = P\left[L_1 = 1 \mid A_0 = 1\right] \neq P\left[L_1 = 1 \mid A_0 = 0\right] = 1 \tag{2.6}$$

Thus, according to standard rules for the estimation of causal effects, *(a)* one cannot adjust for the extraneous risk factor PCP, since it is affected by a previous treatment (AZT); yet one must adjust for PCP since it is a confounder for a later treatment (AP). Thus it may be that, in line with the adage that association need not be causation, neither (2.2) nor (2.5) may represent the direct causal effect of AZT controlling for AP. Since both treatments (AZT and AP) were randomized, one would expect that there should exist a "correct" analysis of the data such that the association observed in the data under that analysis has a causal interpretation as the direct effect of AZT controlling for AP. In the next section, we derive such a "correct" analysis based on the G-computation algorithm of Robins (1986). We show that there is, in fact, no direct causal effect of AZT controlling for AP. That is, given that all subjects take AP, it is immaterial to the survival rate in the study population whether or not AZT is also taken.

3. A Formal Theory of Causal Inference

We use this section to derive the G-computation algorithm formula that will allow a correct analysis of the trial in Section 2.

3.1. The Observed Data

Let $V = \overline{V}_J \equiv (V_0, \ldots, V_J)$ be a $J+1$ random vector of temporally ordered variables with associated distribution function $F_V(v)$ and density $f_V(v)$ with respect to a measure μ where, for any $Z = (Z_0, \ldots, Z_M), \overline{Z}_m \equiv (Z_0, Z_1, \ldots, Z_m)$ is the history through $m, m \leq M$. We assume each component V_j of V is either a real-valued continuous random variable or discrete. The measure μ is the product measure of Lebesgue and counting measures corresponding to the continuous and discrete components of V. Now let $\overline{A}_K = (A_0, A_1, \ldots, A_K)$ be a temporally ordered $K+1$ subvector of V consisting of the treatment variables of interest. Denote by t_k the time at which treatment A_k is received. Let L_k be the vector of all variables whose temporal occurrence is between treatments A_{k-1} and A_k with L_1 being the variables preceding A_1 and L_{K+1} being the variables succeeding A_K. Hence, $\overline{L}_{K+1} = (L_0, \ldots, L_{K+1})$ is the vector of all non-treatment variables. For notational convenience, define $\overline{A}_{-1}, \overline{L}_{-1}, \overline{V}_{-1}$ be identically 0 for all subjects. We view $\overline{A}_K$ as a sequence of treatment (control, exposure) variables whose causal effect on $\overline{L}_{K+1}$ we wish to evaluate. To do so, we must define a feasible treatment regime. Before doing this, in Figure 2, with $K = 1$ we use the trial of Section 2 to clarify our notation.

Figure 2

	V_0	V_1	V_2	V_3	V_4
	L_0	A_0	L_1	A_1	$Y \equiv L_2$
Variable Name:	0	AZT at t_0	PCP at t_1	AP at t_1	survival to t_2

Note that since no data has been collected prior to A_0, L_0 can be set to 0 for all subjects. Also, since all subjects are alive at t_1, we do not require a variable recording survival to t_1.

Remark: Our results do not actually require that the temporal ordering of the univariate components of the vector variable L_k need be known. [Pearl and Robins (1995) discuss conditions under which one may further relax the assumption that the temporal ordering of the variables in V is completely known.]

3.2. Treatment Regimes and Counterfactual Data

We adopt the convention that if two covariate histories, e.g., $\overline{\ell}_m$ and $\overline{\ell}_k$, are used in the same expression with $k > m$ then $\overline{\ell}_m$ is the initial segment of $\overline{\ell}_k$ through t_m. Also $\mathbf{Z}$ will denote the support of a random variable Z. We define a feasible treatment regime g to be a function $g\left(t_k, \overline{\ell}_k\right)$ that assigns to each t_k and $\overline{\ell}_k \in \overline{\mathbf{L}}_k$ a treatment $a_k \in \mathbf{A}_k$. Let $\mathbf{g}$ be the set of all feasible regimes. Given the function $g\left(t_k, \overline{\ell}_k\right)$ of two arguments, we define the function $g\left(\overline{\ell}_k\right)$ of one argument by the relation $g\left(\overline{\ell}_k\right) = \left\{g\left(t_m, \overline{\ell}_m\right); 0 \leq m \leq k\right\}$. Since there is a one-to-one relationship between the functions, we shall identify the regime g with both functions. If $g\left(t_k, \overline{\ell}_k\right)$ is a constant, say a_k^*, not depending on $\overline{\ell}_k$ for each k, we say regime g is non-dynamic and denote it by $g = \overline{a}^*$, $\overline{a}^* = (a_0^*, \ldots, a_K^*)$. Otherwise we say the regime g is dynamic.

Example (3.1): In the hypothetical trial in Sec. 2, the non-dynamic regime in which subjects are forced to take AZT and then AP is $g = (1,1)$. The regime in which subjects are forced not to take AZT but then to take AP is $g = (0,1)$. The dynamic regime g^* in which subjects take AZT and then take AP only if they develop PCP is given by $g^*\left(\overline{\ell}_0\right) = 1$, $g^*\left(\overline{\ell}_1 = 1\right) = (1,1)$ and $g^*\left(\overline{\ell}_1 = 0\right) = (1,0)$.

Remark: In medical trials, subjects are usually assigned to dynamic regimes since, if toxicity develops at t_k (with toxicity recorded in L_k), the treatment must be modified accordingly.

It will now be convenient to consider the realizations of random variables for particular subjects i. We shall say that, in the observed study, subject i has followed a treatment history consistent with a particular regime g through t_k if $g\left(\overline{\ell}_{ki}\right) = \overline{a}_{ki}$.

We reserve the i subscript for study subjects and write $\overline{\ell}_{(K+1)i} \equiv \overline{\ell}_i$ and $\overline{a}_{Ki} \equiv \overline{a}_i$. For each subject i and regime g, we shall assume there exists (possibly counterfactual data) $\overline{\ell}_{(K+1)ig} \equiv \overline{\ell}_{ig}$ and $\overline{a}_{Kig} \equiv \overline{a}_{ig}$ representing the ℓ-history and a-history that would be observed in the closest possible world to this world (Lewis, 1973) in which a subject followed a treatment history consistent with regime g. Here $\overline{\ell}_{ig}$ determines $\overline{a}_{ig}$ through the relationship $g\left(t_k, \overline{\ell}_{kig}\right) = a_{kig}$; this formalizes the idea that subject i would have followed regime g.

Remark (3.1): For certain variables (such as gender) we may not understand what it means to speak of the closest possible world to this one in which one's gender were different. Such variables will not be regarded as potential treatments. (See Robins (1986, 1995b) and Rubin (1978).) In contrast, the literature on causal theories based on *DAGs* (Pearl, 1994; SGS, 1993) appears to suggest that the causal effect of any variable (including gender) is potentially well-defined (but see Heckerman and Shachter, 1995).

We define the counterfactual data on subject i to be $\left\{\overline{\ell}_{ig}, \overline{a}_{ig}; g \in \mathbf{g}\right\}$. Note that, by assumption, the counterfactual data on subject i do not depend on the observed or counterfactual data for any other subject (Rubin, 1978). We shall make a consistency assumption that serves to link the counterfactual data with the observed data. This assumption states that if subject i has an observed treatment history through t_k equal to that prescribed by regime g, then his observed outcome history through t_{k+1} will equal his counterfactual outcome history under regime g through that time.

Consistency Assumption: For $k = 0, \ldots, K$,

$$g\left(\overline{\ell}_{ki}\right) = \overline{a}_{ki} \text{ implies } \overline{\ell}_{(k+1)ig} = \overline{\ell}_{(k+1)i} \tag{3.1}$$

Remark: Interestingly, our main theorems [Theorems (3.1) and (3.2)] only require that (3.1) holds for $k = K$.

We shall assume $\left(\overline{\ell}_i, \overline{a}_i, \left\{\overline{\ell}_{ig}, \overline{a}_{ig}; g \in \mathbf{g}\right\}\right), i = 1, \ldots, n$, are realizations of independent identically distributed random vectors $\left(\overline{L}_i, \overline{A}_i, \left\{\overline{L}_{ig}, \overline{A}_{ig}; g \in \mathbf{g}\right\}\right)$. This will be the case if, as discussed in Robins (1995b), we can regard the n study subjects as randomly sampled without replacement from a large superpopulation of N subjects and our interest is the causal effect of treatment in the superpopulation. For notational convenience, we shall often suppress the i subscript. Robins (1986, 1987) represented the observed data and counterfactual data for the n subjects in "event trees" called causally interpreted structured tree graphs (CISTGs), whose topology is illustrated by Figure 1. These CISTGs are mathematically equivalent to the observed and counterfactual data for the n subjects plus the consistency assumption (3.1). The structure, developed in this section, is sufficient to provide a formal mathematical definition for essentially any English sentence expressing a statement about causal effects. For example, in the context of the trial of Section 2, the null hypothesis that there is no direct effect of AZT on any subject's survival through t_2 controlling for AP (when all subjects receive AP) can, in the notation of Ex. 3.1, be formally expressed as $Y_{ig=(1,1)} = Y_{ig=(0,1)}$ with probability 1.

3.3. Sequential Randomization, the G-computation Algorithm, and a Correct Analysis of the Trial of Sec. 2

We now suppose that, as is usually the case, the outcome of interest is a particular function $Y = y\left(\overline{L}_{K+1}\right)$ of $\overline{L}_{K+1}$ rather than the entire $\overline{L}_{K+1}$. For example, in the trial in Sec. 2, $Y = L_2 \equiv L_{K+1}$ denoted survival through time t_2. The purpose of this section is to give sufficient conditions to identify the law of $Y_g = y\left(\overline{L}_{(K+1)g}\right)$ from the observed data and to provide an explicit computational formula, the G-computation algorithm formula, for the density $f_{Y_g}(y)$ of Y_g under these conditions. If the A_k had been assigned at random by the flip of a coin, then for each regime g we would have, for $k = 0, 1, \ldots, K$, and all $\overline{\ell}_k$

$$\overline{L}_g \coprod A_k \mid \overline{L}_k = \overline{\ell}_k, \overline{A}_{k-1} = g\left(\overline{\ell}_{k-1}\right) \tag{3.2}$$

even if (as in Sec. 2) the probability that the coin landed heads depended on $\overline{L}_k$ and $\overline{A}_{k-1}$. Eq. (3.2) states that among subjects with covariate history $\overline{\ell}_k$ through t_k, and treatment history $g\left(\overline{\ell}_{k-1}\right)$ through t_{k-1}, treatment A_k at time t_k is independent of the counterfactual outcomes $\overline{L}_g \equiv \overline{L}_{K+1,g}$. This is because $\overline{L}_g$, like gender or age at enrollment, is a fixed characteristic of a subject unaffected by the randomized treatments a_k actually received. Note that such physical sequential randomization of the a_k would imply that (3.2) would hold for all $\overline{A}_{k-1}$ histories. However, we shall only require that it hold for the $\overline{A}_{k-1}$ histories $g\left(\overline{L}_{k-1}\right)$ consistent with regime g. Theorem 1 below states that, given (3.1) and (3.2), the density $f_{Y_g}(y)$ is identified provided that the probability density that a subject follows regime g through t_K is non-zero.

Our results also apply to observational studies in which (3.2) is true. However, in observational studies, (3.2) cannot be guaranteed to hold and is not subject to an empirical test. Therefore, it is a primary goal of epidemiologists conducting an observational study to collect in $\overline{L}_k$, data on a sufficient number of covariates to try to make Eq. (3.2) at least approximately true. For example, suppose the data in Figure 1 was from an observational study rather than a sequentially randomized study. Because physicians tend to initiate AP preferentially for subjects who have had a recent bout of PCP, and because recent bouts of PCP signify poor prognosis (i.e., patients with small values of Y_g), Eq. (3.2) would be false if PCP at t_1 were not a component of L_1. It is because physical randomization guarantees Eq. (3.2) that most people accept that valid causal inferences can be obtained from randomized studies (Rubin, 1978).

For any random Z, let $S_Z(z) = P(Z \geq z)$ be the survivor probability where for a j vector Z, we define $Z \geq z \Leftrightarrow (Z_1 \geq z_1, \ldots, Z_j \geq z_j)$. For reasons that will become clear in Sec. 7, it will often be more convenient to state some results in terms of survivor probabilities or cumulative distribution functions [i.e., $1 - S_Z(z)$] than in terms of densities. Adopt the convention that the conditional density and survival probability of an observable random variable given other observable random variables will be denoted by $f(\cdot \mid \cdot)$ and $S(\cdot \mid \cdot)$. The conditional density or survival probability of a counterfactual random variable with respect to regime g given another counterfactual random variable with respect to that regime will be denoted by $f^g(\cdot \mid \cdot)$ or $S^g(\cdot \mid \cdot)$. Finally, the conditional density or survival probability of a counterfactual random variable with respect to regime g given an observed random variable will be denoted by $f^{g0}(\cdot \mid \cdot)$ or $S^{g0}(\cdot \mid \cdot)$. Further we adopt the convention that $g\left(\overline{\ell}_k\right)$ will always represent a realization of the variable $\overline{A}_k$. Thus, for example, $f^{g0}\left(\ell_k \mid \overline{\ell}_{k-1}, g\left(\overline{\ell}_{k-1}\right)\right) \equiv f_{L_{kg} \mid \overline{L}_{k-1}, \overline{A}_{k-1}}\left(\ell_k \mid \overline{\ell}_{k-1}, g\left(\overline{\ell}_{k-1}\right)\right)$, $f\left\{g\left(\overline{\ell}_k\right) \mid \overline{\ell}_k, g\left(\overline{\ell}_{k-1}\right)\right\} \equiv$ $f_{\overline{A}_k \mid \overline{L}_{k-1}, \overline{A}_{k-1}}\left(g\left(\overline{\ell}_k\right) \mid \overline{\ell}_{k-1}, g\left(\overline{\ell}_{k-1}\right)\right)$, $f^g\left(\ell_k \mid \overline{\ell}_{k-1}\right) \equiv f_{L_{kg} \mid \overline{L}_{(k-1)g}}\left(\ell_k \mid \overline{\ell}_{k-1}\right)$.

Using this notation, the condition that for $k = (0, \ldots, K)$

$$f\left\{\overline{\ell}_k, g\left(\overline{\ell}_{k-1}\right)\right\} \neq 0 \Rightarrow f\left\{g\left(\overline{\ell}_k\right) \mid \overline{\ell}_k, g\left(\overline{\ell}_{k-1}\right)\right\} \neq 0 \tag{3.3}$$

states that subjects whose history is consistent with regime g prior to t_k have a positive density of "remaining" on regime g at t_k.

In the following theorem due to Robins (1986, 1987), we adopt the convention that $\overline{\ell}_{-1} \equiv 0, g\left(\overline{\ell}_{-1}\right) \equiv 0$.

Theorem 3.1: If (3.1) - (3.3) hold for regime g, then Eqs. (3.4) - (3.9) are true.

$$f^g\left(\ell_k \mid \overline{\ell}_{k-1}\right) = f^{g0}\left(\ell_k \mid \overline{\ell}_{k-1}, g\left(\overline{\ell}_{k-2}\right)\right) = f\left(\ell_k \mid \overline{\ell}_{k-1}, g\left(\overline{\ell}_{k-1}\right)\right) \tag{3.4}$$

$$S^g\left(y \mid \overline{\ell}_K\right) = S^{go}\left(y \mid \overline{\ell}_K, g\left(\overline{\ell}_{K-1}\right)\right) = S\left(y \mid \overline{\ell}_K, g\left(\overline{\ell}_K\right)\right) \tag{3.5}$$

<u>Remark</u>: Eq. (3.4) says that one time-step ahead innovations in the L-process in the counterfactual world where all subjects followed regime g equal the one step ahead innovations in the observed world among subjects whose previous treatment is consistent with g. Eq. (3.5) states that the conditional distribution of Y given $\overline{L}_K$ in the counterfactual world is equal to the same distribution in the observed world among the subjects whose previous treatment is consistent with g.

$$f^g\left(\bar{\ell}_m \mid \bar{\ell}_k\right) = f^{go}\left(\bar{\ell}_m \mid \bar{\ell}_k, g\left(\bar{\ell}_{k-1}\right)\right) = \prod_{j=k+1}^{m} f\left(\ell_j \mid \bar{\ell}_{j-1}, g\left(\bar{\ell}_{j-1}\right)\right), m > k \tag{3.6}$$

$$f^{g}\left(y,\overline{\ell}_{K}\mid\overline{\ell}_{k}\right)=f^{go}\left(y,\overline{\ell}_{K}\mid\overline{\ell}_{k},g\left(\overline{\ell}_{k-1}\right)\right)=$$
$$f\left(y\mid\overline{\ell}_{K},g\left(\overline{\ell}_{K}\right)\right)\prod_{j=k+1}^{K}f\left(\ell_{j}\mid\overline{\ell}_{j-1},g\left(\overline{\ell}_{j-1}\right)\right) \tag{3.7}$$

Remark: Eq. (3.6) and (3.7) state that multi-step ahead innovations in the counterfactual world can be built up from one step innovations in the observed world. Note (3.6) and (3.7) depend on the law of the observed data only through the densities $f\left(\ell_{j}\mid\overline{\ell}_{j-1},\overline{a}_{j-1}\right)$.

$$S^{g}\left(y\mid\overline{\ell}_{k}\right)=S^{go}\left(y\mid\overline{\ell}_{k},g\left(\overline{\ell}_{k-1}\right)\right)=\int\cdots\iint S\left(y\mid\overline{\ell}_{K},g\left(\overline{\ell}_{K}\right)\right)$$
$$\prod_{j=k+1}^{K}f\left(\ell_{j}\mid\overline{\ell}_{j-1},g\left(\overline{\ell}_{j-1}\right)\right)d\mu\left(\ell_{j}\right) \tag{3.8}$$

$$S^{g}(y)=\int\cdots\iint S\left(y\mid\overline{\ell}_{K},g\left(\overline{\ell}_{K}\right)\right)\prod_{j=0}^{K}f\left(\ell_{j}\mid\overline{\ell}_{j-1},g\left(\overline{\ell}_{j-1}\right)\right)d\mu\left(\ell_{j}\right) \tag{3.9}$$

Remark: Eq. (3.9) states that the marginal survival probability of Y_g is obtained by a weighted average of the $S\left(y\mid\overline{\ell}_{K},g\left(\overline{\ell}_{K}\right)\right)$ with weights proportional to $\omega\left(\overline{\ell}_{K}\right)\equiv\prod_{j=0}^{K}f\left[\ell_{j}\mid\overline{\ell}_{j-1},g\left(\overline{\ell}_{j-1}\right)\right]$. Eq. (3.8) has a similar interpretation except that it conditions on the covariate history $\overline{\ell}_{k}$.

It will be convenient to let $b\left(y,\overline{\ell}_{k},g\right)$ denote the rightmost side of (3.8). Note $b\left(y,\overline{\ell}_{k},g\right)$ depends only on the law of the observables $F_{V}\left(v\right)$, the regime g, y, and $\overline{\ell}_{k}$. In fact, the dependence on $F_{V}\left(v\right)$ is only through the densities $f\left(\ell_{j}\mid\overline{\ell}_{j-1},\overline{a}_{j-1}\right)$ since Y is a function of $\overline{L}_{K+1}$. Similarly, let $b\left(y,g\right)$ denote the right hand side of (3.9).

Example: A Correct Analysis of the Trial of Sec. 2: In the sequential randomized trial of Figure 1, we let $K=1, L_{0}\equiv 0, Y=L_{2}$. Then the probability a subject would survive to $t_{2}\left(Y=1\right)$ if all subjects were treated with AZT at t_{0} and aerosolized pentamidine at t_{1} is $S^{g}\left(y=1\right)$ with g the non-dynamic regime $g=\left(1,1\right)$ and equals, by Eq. (3.9),

$$\sum_{\overline{\ell}_{1}}S\left(y=1\mid\overline{\ell}_{1},g\left(\overline{\ell}_{1}\right)\right)f\left(\overline{\ell}_{1}\mid\ell_{0},g\left(\overline{\ell}_{0}\right)\right)=S\left(y=1\mid\ell_{1}=1,a_{1}=1,a_{0}=1\right)$$
$$f\left(\ell_{1}=1\mid a_{0}=1\right)+S\left(y=1\mid\ell_{1}=0,a_{0}=1,a_{1}=1\right)f\left(\ell_{1}=0\mid a_{0}=1\right)=$$
$$\left(4,000/8,000\right)\left(8,000/16,000\right)+\left(3,000/4,000\right)\left(8,000/16,000\right)=10,000/16,000$$

Similarly $S^{g}\left(y=1\right)$ for regime $g=\left(0,1\right)$ is $10,000/16,000$. Hence, there is, by definition, no direct effect of AZT on survival controlling for AP (when all subjects take AP) since $S^{g=(1,1)}_{(1)}=S^{g=(0,1)}_{(1)}$.

Remark: Note, $f^{g=(1,0)}\left(y=1,\ell_{1}=1\right)$ is not identified, since evaluating (3.7) with $k=0$ we obtain $f\left(y=1\mid\ell_{1}=1,a_{0}=1,a_{1}=0\right)f\left(\ell_{1}=1\mid a_{0}=1\right)$, but the first factor is unidentified on account of the conditioning event having probability zero. This reflects the fact that (3.3) fails, since
$f\left[g\left(t_{1},\overline{\ell}_{1}=1\right)\mid\overline{\ell}_{1}=1,g\left(\overline{\ell}_{0}\right)\right]=f\left[a_{1}=0\mid\ell_{1}=1,a_{0}=1\right]=0$ even though $f\left(\overline{\ell}_{1}=1,g\left(\overline{\ell}_{0}\right)\right)=$
$f\left(\ell_{1}=1,a_{0}=1\right)=\frac{1}{4}\neq 0$.

Remark: If L_k or A_k have continuous components then, in order to avoid measure theoretic difficulties due to the existence of different versions of conditional distributions, we shall impose the assumption that $f\left(\ell_{k}\mid\overline{\ell}_{k-1},\overline{a}_{k-1}\right)$ is continuous in all arguments that represent realizations of random variables with continuous distributions, so that the right hand sides of (3.4)-(3.9) are unique. Gill and Robins (1996) discuss measure theoretic issues in greater depth.

3.4. Sequential Randomization w.r.t. Y Only

Suppose in place of (3.2) we imposed only the weaker condition that, for all k,

$$Y_g \coprod A_k \mid \overline{L}_k = \overline{\ell}_k, \overline{A}_{k-1} = g\left(\overline{\ell}_{k-1}\right). \tag{3.10}$$

When (3.10) holds, we will say that given $\overline{L}$, treatment is (sequentially) randomized with respect to Y_g. When (3.2) holds, we say that, given $\overline{L}$, treatment is (sequentially) randomized w.r.t. regime g. Robins (1992) refers to (3.10) as the assumption of no unmeasured confounders. Robins (1986, 1987) encoded (3.10) and (3.2) in event tree representations called randomized CISTGs w.r.t. Y and randomized CISTGs (RCISTGs). Although (3.10) does not logically imply (3.2), nevertheless, since physical randomization of treatment implies the stronger condition (3.2), one might wonder whether it would ever make substantive sense to impose (3.10) but not (3.2). Interestingly, Robins (1987, 1993) discussed such substantive settings. See Sec. 4 below for an example. For the moment, we restrict ourselves to determining which, if any, parts of Theorem 3.1 remain true under the weaker assumption (3.10) of sequential randomization w.r.t. Y_g.

Theorem 3.2 (Robins, 1987): If (3.1), (3.3), and (3.10) hold for regime g, then Eq. (3.9) and the right-most equalities in Eqs. (3.5) and (3.8) are true. However, $f^g\left(\ell_k \mid \overline{\ell}_{k-1}\right)$, $f^{g0}\left(\ell_k \mid \overline{\ell}_{k-1}, g\left(\overline{\ell}_{k-1}\right)\right)$, and $S^g\left(y \mid \overline{\ell}_k\right)$ are not identified.

Theorem 3.2 states that we can not identify densities that involve the counterfactual random variable $\overline{L}_g$. However, we can identify the marginal distribution of Y_g and the conditional distribution of Y_g given the observed data $\overline{L}_k, \overline{A}_{k-1} = g\left(\overline{L}_{k-1}\right)$. For completeness, a proof of Theorem (3.2) is given in Appendix 1.

4. Confounding

Suppose we believe (3.10) holds for all $g \in \mathbf{g}$ but that a subset $\overline{U}_{K+1}$ of $\overline{L}_{K+1}$ is not observed. The observed subset $\overline{O}_{K+1}$ of $\overline{L}_{K+1}$ includes the outcome variables Y. Define $\overline{U}_k = \overline{L}_k \bigcap \overline{U}_{K+1}$ and $\overline{O}_k = \overline{L}_k \bigcap \overline{O}_{K+1}$ to be the unobserved and observed non-treatment variables through time t_k. The goal of this section is to define restrictions on the joint distribution of $V = \left(\overline{L}_{K+1}, \overline{A}_K\right) \equiv \left(\overline{L}, \overline{A}\right)$ such that (3.10) will imply that for a given $g \in \mathbf{g}\left(O\right) \equiv \left\{g \in \mathbf{g}; g\left(\overline{\ell}_k\right) = g\left(\overline{o}_k\right) \text{ for all } k\right\}$

$$Y_g \coprod A_k \mid \overline{O}_k, \overline{A}_{k-1} = g\left(\overline{O}_{k-1}\right) \text{ for all } k \tag{4.1}$$

since, then, for that $g \in \mathbf{g}\left(O\right)$, by Theorem (3.2), $S^g\left(y\right)$ is identified from data $\left(\overline{A}_K, \overline{O}_{K+1}\right)$ and can be computed by the g-computation algorithm formula (3.9) with o substituted for ℓ. $\mathbf{g}\left(O\right)$ is exactly the set of treatment regimes that a person can follow without having data on $\overline{U}_{K+1}$. Denote by $b\left(y, \overline{\ell}_k, g\right)$ the right hand side of Eq. (3.8). Note $b\left(y, \overline{\ell}_k, g\right)$ is a functional of the law $F_V\left(v\right)$ of V.

Theorem: If, for a regime $g \in \mathbf{g}\left(\mathbf{O}\right)$ and each y, k,

$$E\left[b\left(y, \overline{L}_k, g\right) \mid \overline{O}_k, A_k, \overline{A}_{k-1} = g\left(\overline{O}_{k-1}\right)\right] \\ \textit{does not depend on } A_k \tag{4.2}$$

then Eq. (3.10) implies Eq. (4.1).

Proof: Given (3.10), by Theorem (3.2) we have that Eq. (3.8) implies that (4.2) is equivalent to $E\left[I\left(Y_g > y\right) \mid \overline{O}_k, A_k, \overline{A}_{k-1} = g\left(\overline{O}_{k-1}\right)\right]$ does not depend on A_k, which implies (4.1).

When (3.10) and (4.1) are true, we say that $\overline{U}_{K+1}$ is a non-confounder for the effect of $\overline{A}_K$ on Y given data on $\overline{O}_{K+1}$ (Robins, 1986).

Corollary 4.1: If for regime $g \in \mathbf{g}\left(\mathbf{O}\right)$ and each k, there exists $U_{bk} \subseteq \overline{U}_k$ such that

$$b\left(y, \overline{L}_k, g\right) = b\left(y, \left(\overline{O}_k, U_{bk}\right), g\right) \tag{4.3}$$

and

$$U_{bk} \coprod A_k \mid \overline{O}_k, \overline{A}_{k-1} = g\left(\overline{O}_{k-1}\right), \tag{4.4}$$

then (3.10) implies (4.1).

Proof: The supposition of the Corollary implies Eq. (4.2).

The suppositions of Theorem (4.1) and Corollary (4.1) are difficult to check because $b\left(y, \overline{L}_k, g\right)$ is the high dimensional integral given by the RHS of (3.8). The following Corollary provides restrictions on the law $F_V(v)$ of $V = \left(\overline{L}, \overline{A}\right)$ that can be checked without integration.

Corollary 4.2: If for a regime $g \in \mathbf{g}\left(\mathbf{O}\right)$ and each k, $\overline{U}_k = \left(U_{bk}, U_{ak}\right)$, U_{bk} satisfies (4.4), and $O_k \coprod U_{ak} \mid \overline{O}_{k-1}, \overline{A}_{k-1} = g\left(\overline{O}_{k-1}\right), U_{bk}$, then (3.10) implies (4.1).

Proof: The supposition of the Corollary (4.2) imply Eqs. (4.3) and (4.4).

4.1. Confounding in DAGs

Pearl and Robins (1995) and Robins and Pearl (1996) have recently developed a graphical approach [based on representing $F_V(v)$ by a directed acyclic graph (DAG)] for determining other sufficient non-integral conditions on $F_V(v)$ under which Eq. (3.10) implies (4.1). Specifically, without loss of generality, we represent V by a complete DAG T^* consistent with the ordering of the variables $V = (V_0, \ldots, V_J)$ (Pearl, 1995). That is, V are the vertices (nodes) of T^* and $F_V(v) = \prod_{j=0}^{J} f_{V_j \mid Pa_j}\left(v_j \mid pa_j\right)$ where Pa_j are the parents of V_j on T^* and pa_j and v_j are realizations. By the completeness of the DAG T^*, $Pa_j = \bar{V}_{j-1}$. We can remove arrows from V_m to V_j, $m < j$, on T^* if and only if $f_{V_j \mid \bar{V}_{j-1}}\left(v_j \mid \overline{v}_{j-1}\right)$ does not depend on v_m. Henceforth, we denote by T the DAG in which all such arrows that the investigator knows *a priori* can be removed have been removed. Thus, the resulting DAG T may no longer be complete. Now given a regime $g \in \mathbf{g}\left(O\right)$, let $\overline{Y}^{\dagger}_k$ be the smallest subset of $\overline{O}_k$ such that $g\left(\overline{O}_k\right) = g\left(\overline{Y}^{\dagger}_k\right)$, i.e., the $\bar{Y}^{\dagger}_k$ are the variables actually used to assign treatment A_k at t_k.

Let T^{gm} be the DAG T modified so that for $k \geq m$ *(i)* all directed edges from $\bar{Y}^{\dagger}_k$ to A_k are included and represent the deterministic dependence of $A_k = g\left(\overline{Y}^{\dagger}_k\right)$ on $\overline{Y}^{\dagger}_k$ under regime g, and *(ii)* there are no other edges into A_k. Abbreviate T^{g0} as T^g. We call T^g the g-manipulated graph. T^g is the manipulated graph of SGS (1993). The G-computation algorithm formula $b\left(y, g\right)$ is the marginal survival distribution of Y based on the manipulated graph (Robins, 1995b).

Let T^g_k be the DAG that has edges into $A_1, \ldots, A_k$ as in T, edges into $A_{k+1}, \ldots, A_K$ as in T^g, no edges out of A_k and is elsewhere identical to T. That is, T^g_k is $T^{g(k+1)}$ with all edges out of A_k removed. Suppose Y is a subset of the variables in $\overline{O}_{K+1}$.

We then have

Theorem 4.2: (Robins and Pearl, 1996). If

$$\left(Y \coprod A_k \mid \overline{O}_k, \overline{A}_{k-1}\right)_{T^g_k}, k \leq K, \tag{4.5}$$

then (3.10) implies (4.1). Here $\left(A \coprod B \mid C\right)_{T^g_k}$ stands for d-separation of A and B given C in T^g_k (Pearl, 1995). Note that checking d-separation is a purely graphical (i.e., visual) procedure that avoids any integration.

Theorem 4.3: (Robins and Pearl, 1996). Eq. (4.5) is true if and only if, for $k \leq K$, $\overline{U}_k = (U_{ak}, U_{bk})$ for possibly empty mutually exclusive sets U_{ak}, U_{bk} satisfying *(i)* $\left(U_{bk} \coprod A_k \mid \overline{O}_k, \overline{A}_{k-1}\right)_{T^{g(k+1)}}$ and *(ii)* $\left(U_{ak} \coprod Y \setminus \overline{O}_k \mid \overline{O}_k, \overline{A}_k, U_{bk}\right)_{T^{g(k+1)}}$. Here $A \setminus B$ is the set of elements in A but not in B.

Remark: U_{ak} need not be contained in $U_{a(k+1)}$, and similarly for U_{bk} and $U_{b(k+1)}$.

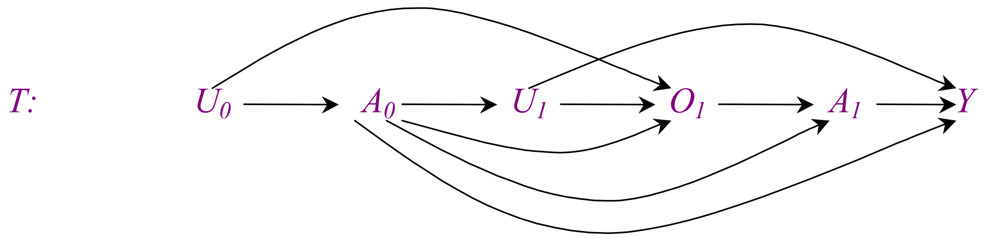
Example (4.1): Suppose underlying the observed variables in Figures 1 and 2 are the variables in Figure 3, where each column represents alternative but equivalent names for a particular variable.

Figure 3

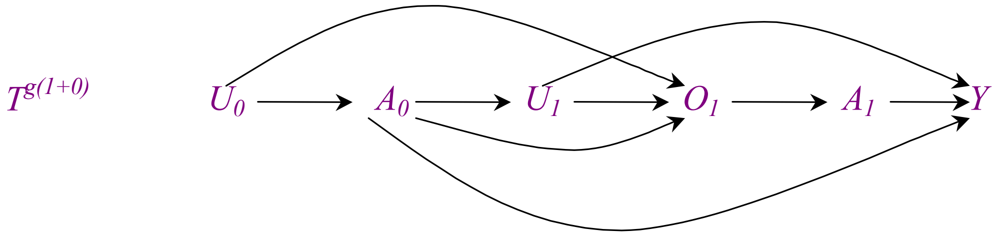
V_0	V_1	V_2	V_3	V_4	V_5
U_0	A_0	U_1	O_1	A_1	$Y = O_2$
L_0	A_0	L_{11}	L_{12}	A_1	L_2
PCP	AZT	Im	PCP	AP	Survival

with $K = 1$ and $L_1 \equiv (L_{11}, L_{12})$ in Fig. 3, $O_1 \equiv L_{12}$ from Fig. 3 is L_1 of Fig. 2, and $O_2 = L_2 = Y$ of Fig. 3 is $L_2 = Y$ of Fig. 2. In Fig. 3, $U_0 = L_0$ is an unmeasured variable that represents whether a subject had PCP prior to the time t_0 of AZT treatment, and $I_m \equiv L_{11}$ is a subject's unrecorded underlying immune status between t_0 and t_1. We now assume that the data in Fig. 1 are from an observational study rather than a sequential randomized trial. Without physical sequential randomization, we can no longer know *a priori* that (3.2) holds for the variables in Figures 1 and 2. However, we shall assume that in our observational study, Eq. (3.2) holds for the variables in Fig. 3. Further, we suppose that the joint distribution of $(V_0, \ldots, V_5)$ can be represented by the DAG T in Fig. 4.

Figure 4



Our goal is to determine whether (4.1) holds for the variables in Fig. 3 (or, equivalently, whether (3.10) holds for the variables in Fig. 2) for the regimes $g = (a_0, a_1) = (1, 1)$ and $g = (0, 1)$. To do so, we use Theorem 4.3.



Hence $T^{g(1+0)}$ satisfies $\left(U_{a0} \coprod Y \mid \overline{A}_0, \overline{O}_0 = \emptyset\right)_{T^{g(1+0)}}$ with $U_{a0} \equiv U_0$, and $U_{b0} \equiv \emptyset$ since the path $U_0 O_1 U_1 Y$ is blocked by the collider O_1. Hence, the hypothesis of Theorem (4.3) holds for $k = 0$. Furthermore, since $T^{g(1+1)} = T$, the hypothesis of Theorem (4.3) holds for $k = 1$ with $U_{b1} = (U_0, U_1), U_{a1} = \emptyset$. Hence, we conclude that, by Theorem (4.3), (4.1) holds for the variables in Fig. 3, so we can estimate the distribution of $Y_{g=(1,1)}$ and $Y_{g=(0,1)}$ from the data (A_0, O_1, A_1, O_2) by the G-computation algorithm. In particular, if the distribution of the observed variables in Fig. 3 (i.e., of the variables in Figs. 1 and 2) is given by Fig. 1, then, by our calculations in Sec. 3, $S^{g=(1,1)}(y = 1) = S^{g=(0,1)}(y = 1) = 10,000/16,000$. [However, neither the hypothesis of Theorem (4.2) nor Eq. (4.1) holds for the regime g^* of Example (3.1) in which subjects take AZT and then take AP only if they develop PCP between t_0 and t_1. In fact, $S^{g^*}(y)$ is not identified.]

Remark: As we have seen in the above example, the graphical conditions in Theorems (4.2) and (4.3) can be checked visually without performing integration. However, they require that we understand enough about the joint law $F_V(v)$ of V so that we can correctly construct an (incomplete) DAG T with appropriate arrows removed from the complete DAG T^*. Since $\bar{U}_{K+1}$ is unobserved, this understanding cannot be based on empirical associations in our data and thus must be based on prior subject matter knowledge or beliefs. In many (but not all) contexts, our prior beliefs concerning causal effects will be sharper than our beliefs concerning the magnitude of non-causal associations (since causal effects tend to be more stable across time and place). Thus, we might have sharper beliefs concerning which arrows can be removed from the complete DAG T^* to form the DAG T if each arrow on T^* represents a direct causal effect. Thus, we might like to include enough variables in $\overline{U}_{K+1}$ and thus in V such that any arrow from V_m to $V_j, j > m$, on T^* represents the causal effect of V_m on V_j controlling for the other variables in $\overline{V}_{j-1}$. Then the absence of an arrow from V_m to V_j on T would represent the absence of a direct causal effect. That is, we would like $f[v_j \mid \overline{v}_{j-1}]$ not to depend on v_m if and only if V_m has no direct effect on V_j controlling for the other variables in $\bar{V}_{j-1}$. This, of course, requires that each V_m can be regarded as a potential treatment which, as discussed in Remark (3.1), may not always be the case.

We will have accomplished this if we include a sufficient number of variables in $V \equiv \overline{V}_J$ such that,

$$\textit{For each } K, K = 0, \ldots, J, \textit{Eq. (3.2) is true with } \overline{A}_K \equiv (V_0, V_1, \ldots, V_{K-1}) \qquad (*)$$

for each $\mathbf{g}$. In that case, V_K is in $\overline{L}_{K+1} \equiv \overline{V}_J / \overline{A}_K$ and, by Theorem (3.1), $f_{V_{K,g=(\overline{v}_{K-1})}}(v_K) \equiv f^{g=(\overline{v}_{K-1})}(v_K) = f(v_K \mid \overline{v}_{K-1})$.

Remark: (*) says that each variable was assigned at random given the past. Robins (1986) encodes assumption (*) in an event tree representation called a full RCISTG (FRCISTG) as fine as V. He shows that (under a completely natural consistency assumption) a FRCISTG as fine as V is an RCISTG (i.e., satisfies (3.2)) for all g when $\overline{A}_K$ now represents any subset of the variables V. A FRCISTG as fine as V (i.e., *) is mathematically equivalent to a causal model (i.e. a non-parametric structural equations model) in the sense of Pearl (1995) and SGS (1993). A DAG T with nodes $V = \overline{V}_J$ is a causal model if $V_K = r_K(Pa_K, \epsilon_K)$ with $K = 0, \ldots, J$, the ϵ_K are mutually independent random variables, and each $r_K(\cdot, \cdot)$ is a deterministic function. To show the equivalence, it is easy to check using the independence of the ϵ_K that a causal model satisfies (*). To see that a FRCISTG as fine as V is a causal model, for $K = 0, \ldots, J$, let $Pa_K = \overline{V}_{K-1}$, define $\epsilon_K = \left\{V_{K,g=\overline{v}_{K-1}}; \overline{v}_{K-1} \in \overline{\mathbf{V}}_{K-1}\right\}$, $r_K(\cdot, \cdot)$ to be the function that selects the appropriate component of ϵ_K, and check that (*) implies mutual independence of the ϵ_K. Using this equivalence, it immediately follows that the manipulation theorem of Spirtes et al. (1993) is a corollary of Theorem (3.1) which was first proved in Robins (1986). See Robins (1995b) for additional discussion.

Example (4.2): Suppose now that, as discussed above, we included the variables U_0 and U_1 in Fig. 3 precisely so that we would believe that condition (*) holds for the entire set of 6 variables. Then (assuming our beliefs are correct) the arrows in DAG T of Fig. 4 have direct causal interpretations. We immediately see from Fig. 4 that, since PCP status prior to time 0 (U_0) determined AZT treatment in A_0, the study cannot possibly represent the sequential randomized trial described in Sec. 2 (since, in that trial, AZT treatment was determined at random by the flip of a fair coin). Further, other than the treatments A_0 and A_1, only the underlying immune status U_1 has a direct causal effect on survival (O_2) since there are no arrows from either U_0 or O_1 to O_2. That is, PCP is not a direct causal risk factor for death controlling for A_0, A_1, and U_1. Further, we observe that the AP treatment (A_1) at t_1 is only influenced by treatment A_0 and measured PCP (O_1). In particular, the decision whether to treat with AP (A_1) at t_1 did not depend on the unmeasured PCP status U_0 or the underlying immune status U_1.

Relationship of Figure 4 to Figure 1:

In the data given in Fig. 1, we have noted in Sec., 2 that PCP status ($L_1 = O_1$) is an independent risk factor for death controlling for the other measured variables A_0 and A_1. From the observed data in Fig. 1 alone, we have no way to determine whether this association is causal or not. However, our assumption ($*$) for the variables in Fig. 3 implies that Fig. 4 represents the underlying causal relationships. Hence the association between PCP and survival controlling for A_0 and A_1 in Fig. 1 represents the fact that the true underlying causal risk factor U_1 of Fig. 4 is not measured and thus, since U_1 causes PCP (O_1), an association between O_1 and survival (conditional on A_0 and A_1) is induced in Fig. 1.

Relationship of Figure 4 to the Causal Hypothesis of Interest:

Our hypothesis of interest, as discussed in Sec. 2, is whether there is a direct causal effect of A_0 on $Y \equiv O_2$ controlling for $A_1 = 1$. Our prior beliefs included in DAG T of Fig. 4 were not sufficient to accept or reject our hypothesis *a priori*. However, our prior beliefs were sufficiently strong that we were able to empirically test our hypothesis by applying the G-computation algorithm to the observed data in Fig. 1. We discovered that there was no direct effect of A_0 on Y controlling for A_1. This absence of a direct effect of A_0 controlling for A_1 could be due to the fact that *(a)* the arrow from A_0 to Y in Fig. 4 is (as a fact of nature) missing and, *(b)* the arrow from A_0 to U_1 is missing. [The arrow from U_1 to Y_2 cannot be missing since, in the data, $Y \not\!\perp\!\!\!\perp O_1 \mid A_0, A_1$.] Alternatively, both these arrows could be present but the magnitude of the effect of A_0 on Y controlling for A_1 represented by the arrow from A_0 to Y is exactly balanced by an equal in magnitude (but opposite in sign) effect of A_0 on Y controlling for A_1 determined by the arrows from A_0 to U_1 and from U_1 to Y. Based solely on our assumptions encoded in DAG T in Fig. 4, the assumption (*), and the data in Fig. 1, we cannot empirically discriminate between these alternative explanations. Note that the direct arrow from A_0 to Y in Fig. 4 represents precisely that part of the possible direct effect of A_0 on Y controlling for A_1 that is not through (mediated by) underlying immune status U_1.

Finally we note that Fig. 4 is an example of the phenomenon discussed in Sec. 3 in which, for the observed data in Fig. 2, (3.2) is false but (3.10) is true for the regime $g = (1, 1)$. That is, for regime $g = (1, 1)$, even if (*) holds for the variables in Fig. 4, we have randomization [in the absence of data on (U_0, U_1)] w.r.t. $Y_g \equiv O_{2g}$ but not w.r.t. O_{1g} and O_{2g} jointly. This is because the association between A_0 (AZT) and O_1 (PCP after t_0) is confounded by the unmeasured factor PCP prior to t_0 (U_0). Thus we can estimate the effect of regime $g = (1, 1)$ on survival but not its effect on PCP at t_1 (O_1); as a consequence, as noted above, we cannot estimate the effect of the dynamic regime g^* of Example (3.1) on survival [i.e., Eq. (3.10) is false for Y_{g^*}].

5. The g-Null Hypothesis

In many settings, the treatments $\overline{A}_K = (A_0, \ldots, A_K)$ represent a single type of treatment given at different times. For example, all the A_k may represent AP doses at various times t_k. Often, in such settings, an important question is whether the g-null hypothesis of no effect of treatment on Y is true, i.e.,

g-Null Hypothesis:

$$S^{g_1}(y) = S^{g_2}(y) \text{ for all } y \text{ and } g_1, g_2 \in \mathbf{g} \ . \tag{5.1}$$

This reflects the fact that if the **g**-null hypothesis is true, the distribution of Y will be the same under any choice of regime g and thus it does not matter whether the treatment is given or withheld at any time. Now when (3.1), (3.3), and (3.10) are true for all $g \in \mathbf{g}$, Theorem (3.2) implies that the **g**-null hypothesis is true if and only if the following "**g**"-null hypothesis is true.

"g"-Null Hypothesis:

$$b(y, g_1) = b(y, g_2) \text{ for all } y \text{ and } g_1, g_2 \in \mathbf{g} \tag{5.2}$$

where $b(y, g)$ is given by the rightmost side of Eq. (3.9) and depends only on the joint distribution $F_V(v)$ of the observables.

Remark 5a: The "**g**"-null hypothesis is not implied by the weaker condition that for all $g \in \mathbf{g}$, (3.1), (3.3), and (3.10) are true, and $b(y, g = (\overline{a}_1)) = b(y, g = (\overline{a}_2))$ for all non-dynamic regimes $\overline{a}_1 \equiv \overline{a}_{1K}$ and $\overline{a}_2 \equiv \overline{a}_{2K}$. However, the following Lemma is true.

Lemma: The "**g**"-null hypothesis is true if and only if

$$b\left(y, \overline{\ell}_k, g = (\overline{a}_1)\right) = b\left(y, \overline{\ell}_k, g = (\overline{a}_2)\right) \text{ for each } y, k, \overline{\ell}_k, \overline{a}_1, \overline{a}_2 \text{ with } \overline{a}_1 \text{ and } \overline{a}_2$$
$$\text{agreeing through } t_{k-1}, \text{ i.e., } \overline{a}_{1(k-1)} = \overline{a}_{2(k-1)}.$$

Corollary: If (3.1), (3.3), and (3.10) hold for all non-dynamic regimes $g = (\overline{a})$, and, for each $y, \overline{\ell}_k, \overline{a}_{k-1}$, $S^{g_1 0}\left(y \mid \overline{\ell}_k, \overline{a}_{k-1}\right) = S^{g_2 0}\left(y \mid \overline{\ell}_k, \overline{a}_{k-1}\right)$ for any $g_1 = (\overline{a}_1)$, and $g_2 = (\overline{a}_2)$ with $\overline{a}_{1(k-1)} = \overline{a}_{2(k-1)} = \overline{a}_{(k-1)}$, then the "**g**"-null hypothesis holds. The usefulness of this Corollary is that we only require randomization with respect to Y_g for non-dynamic regimes.

Remark 5b: The above results can be generalized to the case where (3.3) may not hold for all $g \in \mathbf{g}$. Let $\mathbf{g}^*$ be the subset of $\mathbf{g}$ on which $b(y, g)$ is defined. $\mathbf{g}^*$ may be strictly contained in $\mathbf{g}$ if (3.3) is not always true. Define the $\mathbf{g}^*$-null hypothesis and the "$\mathbf{g}^*$"-null hypothesis as in (5.1) and (5.2), except with $\mathbf{g}^*$ replacing $\mathbf{g}$. Then we have that the $\mathbf{g}^*$-null hypothesis is true if and only if the "$\mathbf{g}^*$"-null hypothesis is true.

Remark 5c: Difficulties in testing the "$\mathbf{g}^*$"-null hypothesis: The following difficulties arise when attempting to construct tests of the "**g**"-null hypothesis or the "$\mathbf{g}^*$"-null hypothesis.

First each $b(y, g)$ is a high dimensional integral that cannot be computed analytically and thus must be evaluated by a Monte Carlo approximation - the Monte Carlo G-computation algorithm described by Robins (1987, 1989). Second the cardinality of **g** is enormous. However, as we now discuss, the greatest difficulty is attributable to the fact that the "**g**"-null hypothesis does not imply either that the component $f\left(\ell_j \mid \overline{\ell}_{j-1}, \overline{a}_{j-1}\right)$ of $b(y, g)$ does not depend on $\overline{a}_{j-1}$ or that the component $S\left(y \mid \overline{\ell}_K, \overline{a}_K\right)$ does not depend on $\bar{a}_K$. Rather, it only implies that the entire integral $b(y, g)$ does not depend on g. In many settings where the "g"-null or "$\mathbf{g}^*$"-null hypothesis holds, each of the above components of $b(y, g)$ will depend on the treatment history.

Example (5.1): In a modified Fig. 1 in which the survival rate in the 4,000 subjects with history $(A_0 = 1, L_1 = 0, A_1 = 0)$ is .75 rather than .25, the "**g***-null" hypothesis is true [since $b(y, g)$ equal 10,000/16,000 for all three g for which $b(y, g)$ can be computed, i.e., the regimes $g = (1, 1)$, $g = (0, 1)$, and g^* of example (3.1)]. Yet both of the above components of $b(y, g)$ depend on treatment history $\overline{a}$. Further, we have seen in Example (4.2) that the data in Fig. 1 or the modified version of Fig. 1 could be obtained from a fairly realistic setting in which the underlying causal theory is given by DAG T of Fig. 4 under assumption (*).

This example is a particular case of the following general result.

Lemma: If *(i)* $Y = L_{K+1}$; *(ii)* $\overline{L}_K$ is an independent predictor of Y, i.e. $f\left(y \mid \overline{\ell}_K, \overline{a}_K\right)$ depends on $\overline{\ell}_K$; *(iii)* $\overline{A}_{j-1}$ is the independent predictor of L_j, i.e., $f\left(\ell_j \mid \overline{\ell}_{j-1}, \overline{a}_{j-1}\right)$ depends on $\overline{a}_{j-1}$; *(iv)* there is no effect of treatment on the outcome Y so the **g**-null hypothesis holds; and *(v)* (3.1), (3.3), and (3.10) hold as in a sequential randomized trial, so the "**g**"-null hypothesis is true; then $f\left(y \mid \overline{\ell}_K, \overline{a}_K\right)$ must also depend on $\overline{a}_K$. The suppositions of the Lemma will hold in a sequential randomized trial in which the time-varying treatment has no effect on the outcome of interest Y, but the treatment does effect an intermediate outcome that (like PCP) *(i)* is not itself a direct cause of Y, but *(ii)* is caused by and thus correlated with an unmeasured direct causal risk factor for Y (such as immune status).

When both $f\left(y \mid \overline{\ell}_K, \overline{a}_K\right)$ and $f\left(\ell_j \mid \overline{\ell}_{j-1}, \overline{a}_{j-1}\right)$ depend on a-history and ℓ_j is multivariate with continuous components (so, given realistic sample sizes, parametric or semiparametric models are required), the most straightforward approach to testing the "**g**"-null hypothesis will fail.

Specifically, the most straighforward parametric likelihood-based approach to estimation of $b(y, g)$ is to specify parametric models $f\left(\ell_j \mid \overline{\ell}_{j-1}, \overline{a}_{j-1}; \Delta\right)$ and $S\left(y \mid \overline{\ell}_K, \overline{a}_K; \theta\right)$. This parameterization corresponds to the standard parameterization of a DAG model, since we are parameterizing the conditional distribution of each variable given its parents. The maximum likelihood estimators $\widehat{\Delta}$ and $\widehat{\theta}$ of Δ and θ are obtained by maximizing $\prod_i f\left(Y_i \mid \overline{L}_{Ki}, \overline{A}_{Ki}; \theta\right) \prod_{j=0}^{K} f\left(L_{ji} \mid \overline{L}_{(j-1)i}, \overline{A}_{(j-1)i}; \Delta\right)$. $b(y, g)$ is then estimated (using Monte Carlo integration) with model derived estimators substituted for the unknown conditional laws in the rightmost side of (3.9). Robins (1986) provided a worked examples of this approach in the context of a survival outcome Y. However, it is clear from inspecting the right-hand side of (3.9) that there will in general be no parametric subvector, say ψ, of (θ, Δ) that will take a fixed value, say 0, whenever the "g"-null hypothesis (5.2) holds. In particular, this problem will arise whenever, as in Example 5.1, the "g"-null hypothesis (5.2) holds but both $f\left(\ell_j \mid \overline{\ell}_{j-1}, \overline{a}_{j-1}; \Delta\right)$ and $S\left(y \mid \overline{\ell}_K, \overline{a}_K; \theta\right)$ depend on $\overline{a}$ history. In fact, I believe that, for non-linear models, it is essentially impossible to specify an unsaturated parametric or semiparametric model for which there exists a parameter vector, (Δ^*, θ^*) say, such that $f\left(\ell_j \mid \overline{\ell}_{j-1}, \overline{a}_{j-1}; \Delta^*\right)$ depends on $\overline{a}_{j-1}$ and $f\left(y \mid \overline{\ell}_K, \overline{a}_K; \theta^*\right)$ depends on both $\overline{\ell}_K$ and $\overline{a}_K$ and yet the "**g**"-null hypothesis holds. As a consequence, in settings like that of Example (5.1), the "**g**"-null hypothesis, even when true, will almost certainly be rejected whatever data are obtained! Specifically, suppose each L_k is discrete with p-levels, each A_k is discrete with p^*-levels and $Y = L_{K+1}$. Then it follows from Eq. (6.1) below that the "g"-null hypothesis implies $\Omega \equiv (p^* - 1)(p - 1)\left\{(p^* p)^{K+1} - 1\right\} p / (p^* p - 1)$ independent equality constraints on the joint distribution of the observables so, in general, most non-linear models with fewer than Ω parameters will reject the "g"-null hypothesis before any data are obtained. If either L_k or A_k has continuous components, an uncountable number of equality constraints must hold and the above difficulty will be even worse. In contrast, linear models need not imply pre-data rejection of the "g"-null hypothesis. For example, suppose $V = \left(\overline{A}_1, \overline{L}_2\right)$, $Y = L_2$, $L_0 \equiv 0$ and all variables are continuous. Suppose we specify the linear models $Y \mid \overline{A}_1, \overline{L}_1 \sim N\left(\beta_0 + \beta_1 A_0 + \beta_2 A_1 + \beta_3 L_1, \sigma_y^2\right)$, $L_1 \mid A_0 \sim N\left(\alpha_0 + \alpha_1 A_0, \sigma_{\ell_1}^2\right)$. Then a sufficient condition for the "g"-null hypothesis to hold is easily shown to be $\beta_2 = 0$ and $\beta_1 + \beta_3\alpha_1 = 0$. The above difficulty in testing the "**g**"-null hypothesis under the standard DAG parameterization motivated the development of the structural nested models described in Sec. 7. Further motivation for these models is now provided by describing a proper approach to testing the "**g**"-null hypothesis.

6. g-Null Theorem and g-Null Tests

A proper approach to testing the "g"-Null Hypothesis (5.2) is based on the following g-null theorem:

g-null theorem: The "**g**"-Null Hypothesis (5.2) is true $\Leftrightarrow$

$$Y \coprod A_k \mid \overline{L}_k, \overline{A}_{k-1}, \quad k = (0, \ldots, K) \tag{6.1}$$

<u>Example</u>: In the modified version of Fig. 1 described in Example (5.1),

$$f\,(y = 1 \mid a_0 = 0) = f\,(y = 1 \mid a_0 = 1) = 10/16$$

and

$$f\,(y = 1 \mid a_0 = 1, \ell_1 = 0, a_1 = 1) = f\,(y = 1 \mid a_0 = 1, \ell_1 = 0, a_1 = 0) = 3/4 \tag{6.2}$$

so (6.1) is true (as it must be) since we showed in Example (5.1) above that (5.2) is true. However, in the unmodified Fig. 1, (6.2) is false, which implies (5.2) must be false; in particular, $10/16 = b\,(y, g = (1,1)) = b\,(y, g = (0,1)) \neq b\,(y, g^*)$ with regime g^* as defined in Example (3.1).

<u>**Remark**</u>: If L_k is discrete with p-levels, A_k is discrete with p^*-levels, and $Y = L_{K+1}$, by writing (6.1) as $f\left(A_k \mid \overline{L}_k, \overline{A}_{k-1}, Y\right) = f\left(A_k \mid \overline{L}_k, \overline{A}_{k-1}\right)$, $k = (0, \ldots, K)$, it is straight- forward to show that the joint distribution of the observables satisfies
$(p-1)\,(p^*-1)\left\{(p^*p)^{K+1} - 1\right\} p/\,(p^*p - 1)$ independent equality constraints under the "g"-null hypothesis.

<u>**Remark**</u>: The g-null theorem may appear less surprising upon realizing that, given (3.10) is true for all $g \in \mathbf{g}$, the following sharp null hypothesis implies both (5.2) and (6.1).

$$\begin{gathered} \textit{Sharp g-null hypothesis} : Y = Y_{g_1} = Y_{g_2} \textit{ with} \\ \textit{probability 1 for all } g_1, g_2 \in \mathbf{g} \end{gathered} \tag{6.3}$$

Specifically (3.10) plus (6.3) implies (6.1). Further, (6.3) implies the g-null hypothesis (5.1) which, given (3.10), implies (5.2).

We use (6.1) to construct tests of the "g"-null hypothesis. Consider first the case in which A_k is dichotomous, Y is univariate, and L_k is discrete with a small number of levels. Let $Num\left(\overline{\ell}_k, \overline{a}_{k-1}\right)$ and $V\left(\overline{\ell}_k, \overline{a}_{k-1}\right)$ be the numerator and its estimated variance of a two sample test (e.g., t-test, log-rank or Mantel extension test, Wilcoxon test) of the hypothesis that subjects with history $a_k = 1, \overline{\ell}_k, \overline{a}_{k-1}$ have the same distribution of Y as subjects with history $a_k = 0, \overline{\ell}_k, \overline{a}_{k-1}$, satisfying, under (6.1), $E\left[Num\left(\overline{\ell}_k, \overline{a}_{k-1}\right)\right] = 0$ and $E\left\{V\left(\overline{\ell}_k, \overline{a}_{k-1}\right)\right\} = Var\left\{Num\left(\overline{\ell}_k, \overline{a}_{k-1}\right)\right\}$. This relationship will be satisfied for most two sample tests. Let $Num = \sum\limits_{k=0}^{K} \sum_{\overline{\ell}_k} \sum_{\overline{a}_{k-1}} Num\left(\overline{\ell}_k, \overline{a}_{k-1}\right)$ and $V =$
$\sum\limits_{k=0}^{K} \sum_{\overline{\ell}_k} \sum_{\overline{a}_{k-1}} V\left(\overline{\ell}_k, \overline{a}_{k-1}\right)$. It can be shown that, under (6.1), the terms in Num are all uncorrelated (Robins, 1986). Thus, under regularity conditions, $Num/\sqrt{V}$ will have an asymptotic standard normal distribution (Robins, 1986) and will thus constitute a non-parametric test of the "g"-null hypothesis (5.2).

This will be a non-parametric test of the g-null hypothesis (5.1) if Eq. (3.10) holds for all $g \in \mathbf{g}$. However, the power of this test may be poor. This reflects the fact that for k greater than 2 or 3, $Num\left(\overline{\ell}_k, \overline{a}_{k-1}\right)$ will usually be zero because it would be rare (with the sample sizes occurring in practice) for 2 subjects to have the same $\overline{\ell}_k, \overline{a}_{k-1}$ history but differ on a_k.

To obtain greater power for discrete a_k and ℓ_k and to handle a_k and ℓ_k with continuous components, tests of (6.1) can be based on parametric or semiparametric models. For instance, suppose A_k is dichotomous and we can correctly specify a logistic model

$$f\left(A_m = 1 \mid \overline{L}_m, \overline{A}_{m-1}\right) = \left\{1 + \exp\left(-\alpha_0' W_m\right)\right\}^{-1} \tag{6.4}$$

where W_m is a p-dimensional function of $\overline{L}_m, \overline{A}_{m-1}$.

Let $Q_m \equiv q\left(Y, \overline{L}_m, \overline{A}_{m-1}\right)$ where $q(\cdot, \cdot, \cdot)$ is any known real-valued function chosen by the data analyst. Let θ be the coefficient of Q_m when θQ_m is added to the regressors $\alpha_0' W_m$ in (6.4). If, for each m, (6.4) is true for some α_0, then hypothesis (6.1) is equivalent to the hypothesis the true value θ_0 of θ is zero. A score test of the hypothesis $\theta_0 = 0$ can then be computed using logistic regression software where, when fitting the logistic regression model, each subject is regarded as contributing $K+1$ independent observations - one at each treatment time $t_0, t_1, \ldots, t_K$. Robins (1992) provides mathematical justification. This is a semiparametric test since it only requires we specify a parametric model for a L.H.S. of (6.4) rather than for the entire joint distribution of V. As discussed later, the choice of the function $q(\cdot, \cdot, \cdot)$ will effect the power of the test but not its α-level.

Robins (1992) refers to these tests as the g-tests. They are extensions of the propensity score methods of Rosenbaum and Rubin (1983) and Rosenbaum (1984) to time-dependent treatments and covariates.

7. Structural Nested Models:

In this Section we motivate, define, and characterize the class of structural nested models. In this Section, we shall only consider the simplest structural nested model - a structural nested distribution model for a univariate continuous outcome measured after the final treatment time t_K. Robins (1989, 1992, 1994, 1995) considers generalizations to discrete outcomes, multivariate outcomes, and failure time outcomes. Discrete outcomes are also treated in Sec. 8 below.

7.1. Motivation for Structural Nested Models

We assume L_{K+1} is a univariate continuous-valued random variable with a continuous distribution function and denote it by Y. To motivate SNMs, we first note that none of the tests of the "g"-null hypothesis discussed in Sec. 6 was a parametric likelihood-based test. Further, our g-test of the "g"-null hypothesis is unlinked to any estimator of $b(y, g)$ based on the G-computation algorithm. Our first goal in this subsection will be to derive a complete reparameterization of the joint distribution of $V \equiv \left(\overline{L}, \overline{A}\right)$ that will offer a unified likelihood-based approach to testing the "g"-null hypothesis (5.2) and estimating the function $b(y, g)$. Then in Sec. 8 we will develop a unified approach to testing (5.2) and estimating $b(y, g)$ based on the semiparametric g-test of Sec. 6.

7.2. New Characterizations of the "g"-Null and g-Null Hypotheses

Denote by $b\left(y, \overline{\ell}_k, g, F\right)$ the right-hand side of (3.8) where the dependence on the law F of V is made explicit. The first step in constructing a reparameterization of the likelihood is a new characterization of the "g"-null hypothesis (5.2). We assume the conditional distribution of Y given$\left(\overline{\ell}_m, \overline{a}_m\right)$ has a continuous density with respect to Lebesgue measure. Given any treatment history $\overline{a} = \overline{a}_K$, adopt the convention that $(\overline{a}_m, 0)$ will be the treatment history $\overline{a}_K^{(1)}$ characterized by $a_k^{(1)} = a_k$ if $k \leq m$ and $a_k^{(1)} = 0$ if $k > m$.

Let $\gamma\left(y, \overline{\ell}_m, \overline{a}_m, F\right)$ be the unique function satisfying

$$b\left[\gamma\left(y, \overline{\ell}_m, \overline{a}_m, F\right), \overline{\ell}_m, g = \left(\overline{a}_{m-1}, 0\right), F\right] = b\left(y, \overline{\ell}_m, g = \left(\overline{a}_m, 0\right), F\right) \tag{7.1}$$

It follows from its definition that: *(a)* $\gamma\left(y, \overline{\ell}_m, \overline{a}_m, F\right) = y$ if $a_m = 0$; *(b)* $\gamma\left(y, \overline{\ell}_m, \overline{a}_m, F\right)$ is increasing in y; and *(c)* the derivative of $\gamma\left(y, \overline{\ell}_m, \overline{a}_m, F\right)$ w.r.t. y is continuous. Examples of such functions are

$$\gamma\left(y, \overline{\ell}_m, \overline{a}_m, F\right) = y + 2a_m + 3a_m a_{m-1} + 4a_m w_m \tag{7.2}$$

where w_m is a given univariate function of $\overline{\ell}_m$ and

$$\gamma\left(y, \overline{\ell}_m, \overline{a}_m, F\right) = y \exp\left\{2a_m + 3a_m a_{m-1} + 4a_m w_m\right\} \tag{7.3}$$

Our interest in $\gamma\left(y, \overline{\ell}_m, \overline{a}_m, F\right)$ is based on the following theorem proved in Robins (1989, 1995a).

Theorem (7.1): $\gamma\left(y, \overline{\ell}_m, \overline{a}_m, F\right) = y$ for all $y, m, \overline{\ell}_m, \overline{a}_m$ if and only if the "g"-null hypothesis (5.2) holds.

Interpretation in terms of counterfactuals: Although Theorem (7.1) is a theorem referring only to the joint distribution of the observables, it is of interest to examine its implications for the distribution of the counterfactual variables under (3.1) and (3.10). Let $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right)$ be the unique function such that, conditional on the observed data $\overline{\ell}_m, \overline{a}_m$, $\gamma^{\dagger}\left(Y_{g=(\overline{a}_m,0)}, \overline{\ell}_m, \overline{a}_m\right)$ has the same distribution as $Y_{g=(\overline{a}_{m-1},0)}$. If conditional on $\overline{\ell}_m, \overline{a}_m$, $\gamma^{\dagger}\left(Y_{g=(\overline{a}_m,0)}, \overline{\ell}_m, \overline{a}_m\right) = Y_{g=(\overline{a}_{m-1},0)}$ with probability 1, we say that we have local rank preservation since then, conditionally, $y_{ig=(\overline{a}_m,0)} > y_{jg=(\overline{a}_m,0)} \Leftrightarrow y_{ig=(\overline{a}_{m-1},0)} > y_{jg=(\overline{a}_{m-1},0)}$ so the ranks (orderings) of any two subjects i's and j's Y values will be preserved under these alternative treatment regimes. It is clear that, if we have local rank preservation, $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right)$ applied to $Y_{g=(\overline{a}_m,0)}$ removes the effect on Y of a final brief blip of treatment of magnitude a_m at t_m leaving us with $Y_{g=(\overline{a}_{m-1},0)}$. We thus call $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right)$ the "blip-down" function. Local rank preservation is a strong untestable assumption that we would rarely expect to hold, but even without local rank preservation we still view $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right)$ as representing the effect of a final blip of treatment of size a_m since its maps (conditional on $\overline{\ell}_m, \overline{a}_m$) quantiles of $Y_{g=(\overline{a}_m,0)}$ into those of $Y_{g=(\overline{a}_{m-1},0)}$: from its definition $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right)$ satisfies

$$S^{g=(\overline{a}_{m-1},0),0}\left\{\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right) \mid \overline{\ell}_m, \overline{a}_m\right\} = S^{g=(\overline{a}_m,0),0}\left(y \mid \overline{\ell}_m, \overline{a}_m\right) \tag{7.4}$$

But, given (3.1), (3.3), and the sequential randomization assumption (3.10) hold for all $g \in \mathbf{g}$, we have that Theorem (3.2) and Eq. (3.8) imply $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right) = \gamma\left(y, \overline{\ell}_m, \overline{a}_m\right)$. Hence, Theorem (7.1) states:

> Given (3.1), (3.3), and (3.10) hold for all $g \in \mathbf{g}$, there is no effect of a final blip of treatment of size a_m among subjects with each observed history $\overline{a}_m, \overline{\ell}_m$ [i.e., $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right) = y$] if and only if the g-null hypothesis (5.1) is true.

Remark: If the randomization assumption (3.10) is false, $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right) = y$ does not imply the g-null hypothesis: for example, if $K = 2$ then, for subjects with observed history $(A_0 = 0, A_1 = 1)$, there may be an effect on $Y = L_2$ of the treatment a_1 when following the non-dynamic regime $g = (a_0 = 1, a_1 = 1)$.

Remark: If (3.2) is also true for all $g \in \mathbf{g}$, it follows from Theorem (3.1) and Eq. (3.8) that $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right)$ also satisfies the following relationship among strictly counterfactual variables.

$$S^{g=(\overline{a}_{m-1},0)}\left\{\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right) \mid \overline{\ell}_m\right\} = S^{g=(\overline{a}_m,0)}\left(y \mid \overline{\ell}_m\right) \tag{7.5}$$

7.3. Pseudo-Structural and Structural Nested Distribution Models

In view of theorem (7.1), our approach will be to construct a parametric model for $\gamma(y, \overline{\ell}_m, \overline{a}_m, F)$ depending on a parameter ψ such that $\gamma(y, \overline{\ell}_m, \overline{a}_m, F) = y$ if and only if the true value ψ_0 of the parameter is 0. We will then reparameterize the likelihood of the observables $(\overline{L}, \overline{A})$ in terms of a random variable which is a function of the observables and the function $\gamma(y, \overline{\ell}_m, \overline{a}_m, F)$. As a consequence, likelihood-based tests of the hypothesis $\psi_0 = 0$ will produce likelihood-based tests of the "g"-null hypothesis (5.2).

Definition: The distribution $F \equiv F_V$ of $V \equiv \left(\overline{L}, \overline{A}\right)$ follows a pseudo-structural nested distribution model (PSNDM) $\gamma^* \left(y, \overline{\ell}_m, \overline{a}_m, \psi_0\right)$ if $\gamma \left(y, \overline{\ell}_m, \overline{a}_m, F\right) = \gamma^*(y, \overline{\ell}_m, \overline{a}_m, \psi_0)$ where $\gamma^*(\cdot, \cdot, \cdot, \cdot)$ is a known function; (2) ψ_0 is a finite vector of unknown parameters to be estimated taking values in R^p; (3) for each value of $\psi \epsilon R^p$, $\gamma^* \left(y, \overline{\ell}_m, \overline{a}_m, \psi\right)$ satisfies the conditions *(a)*, *(b)*, *(c)* and *(d)* that were satisfied by $\gamma \left(y, \overline{\ell}_m, \overline{a}_m, F\right)$; (4) $\partial\gamma^* \left(y, \overline{\ell}_m, \overline{a}_m, \psi\right) / \partial\psi'$ and $\partial^2\gamma^* \left(y, \overline{\ell}_m, \overline{a}_m, \psi\right) / \partial\psi' \partial t$ are continuous for all ψ; and (5) $\gamma^* \left(y, \overline{\ell}_m, \overline{a}_m, \psi\right) = y$ if $\psi = 0$ so that $\psi_0 = 0$ represents the "g-" null hypothesis (5.2).

We say the law of $\left(\overline{L}, \overline{A}, \left\{Y_g, \overline{L}_g, \overline{A}_g; g \in \mathbf{g}\right\}\right)$ follows a structural nested distribution model (SNDM) $\gamma^* \left(y, \overline{\ell}_m, \overline{a}_m, \psi_0\right)$ if the definition of a pseudo-structural nested model holds when we replace $\gamma \left(y, \overline{\ell}_m, \overline{a}_m, F\right)$ by $\gamma^\dagger \left(y, \overline{\ell}_m, \overline{a}_m\right)$. Theorem (3.2) implies that, given (3.1), (3.3), and (3.10) hold for all $g \in \mathbf{g}$, the function $\gamma^* \left(y, \overline{\ell}_m, \overline{a}_m, \psi_0\right)$ is a structural nested model if and only if it is a pseudo-structural nested model.

Examples of appropriate functions $\gamma^* \left(y, \overline{\ell}_m, \overline{a}_m, \psi\right)$ can be obtained from Eqs. (7.2) and (7.3) by replacing the quantities 2, 3 and 4 by the components of $\psi = (\psi_1, \psi_2, \psi_3)$. We call models for $\gamma \left(y, \overline{\ell}_m, \overline{a}_m, F\right)$ pseudo-structural because, although they mimic structural nested models, pseudo-SNDM are models for the observable rather than the counter-factual random variables.

7.4. The Reparameterization of the Likelihood

Next we recursively define random variables $H_K(\gamma), \ldots, H_0(\gamma)$ that depend on the observables $V \equiv \left(\overline{L}, \overline{A}\right)$ as follows. $H_K(\gamma) \equiv \gamma\left(Y, \overline{L}_K, \overline{A}_K, F\right)$, $H_m(\gamma) \equiv h_m\left(Y, \overline{L}_K, \overline{A}_K, \gamma\right) \equiv \gamma\left(H_{m+1}(\gamma), \overline{L}_m, \overline{A}_m, F\right)$, and $H(\gamma) \equiv H_0(\gamma)$. Note by Theorem (7.1) if the "g"-null hypothesis (5.2) is true, then $H(\gamma) = Y$.

Example: If $\gamma \left(y, \overline{\ell}_m, \overline{a}_m, F\right)$ is given by Eq. (7.2), then $H_m(\gamma) = Y + \sum\limits_{k=m}^{K} (2A_k + 3A_kA_{k-1} + 4A_kW_k)$.

Interpretation in terms of Counterfactuals: Let $H\left(\gamma^\dagger\right)$ be defined like $H(\gamma)$ but with the (counterfactual) blip-down function $\gamma^\dagger$ replacing the function γ. Robins (1993) proved that $H\left(\gamma^\dagger\right)$ has, under consistency assumption (3.1), the same marginal distribution as $Y_{g=(0)}$, the counterfactual variable when treatment is always withheld. Following Robins et al. (1992), we say $H\left(\gamma^\dagger\right)$ is the value of Y when "blipped all the way down."

Since $\gamma\left(Y, \overline{L}_m, \overline{A}_m, F\right)$ is increasing in Y, the map from $V \equiv \left(Y, \overline{L}_K, \overline{A}_K\right)$ to $\left(H(\gamma), \overline{L}_K, \overline{A}_K\right)$ is 1-1 with a strictly positive Jacobian determinant. Therefore, $f_{Y,\overline{L}_K,\overline{A}_k}\left(Y, \overline{L}_K, \overline{A}_K\right) = \{\partial H(\gamma)/\partial Y\} f_{H(\gamma),\overline{L}_K,\overline{A}_K}\left(H(\gamma), \overline{L}_K, \overline{A}_K\right)$. However, Robins (1989, 1995) proves that

$$A_m \coprod H(\gamma) \mid \overline{L}_m, \overline{A}_{m-1} \tag{7.6}$$

It follows that

$$f_{Y,\overline{L}_K,\overline{A}_K}\left(Y, \overline{L}_K, \overline{A}_K\right) = \{\partial H(\gamma)/\partial Y\} f\{H(\gamma)\}$$

$$\prod_{m=0}^{K} f\left(L_m \mid \overline{L}_{m-1}, \overline{A}_{m-1}, H(\gamma)\right) f\left[A_m \mid \bar{L}_m, \bar{A}_{m-1}\right] \tag{7.7}$$

(7.7) is the aforementioned reparameterization of the density of the observables.

Thus we have succeeded in reparameterizing the joint density of the observables in terms of the function $\gamma\left(y,\overline{\ell}_m,\overline{a}_m,F\right)$, its derivative with respect to y, and the densities $f\left[\ell_m \mid \overline{\ell}_{m-1},\overline{a}_{m-1},h\left(\gamma\right)\right]$, $f\left(h\left(\gamma\right)\right)$, and $f\left[a_m \mid \overline{\ell}_m,\overline{a}_{m-1}\right]$.

We can then specify a fully parametric model for the joint distribution of the observables by specifying *(a)* a pseudo-SNFTM $\gamma^*\left(y,\overline{\ell}_m,\overline{a}_m,\psi_0\right)$ for $\gamma(y,\overline{\ell}_m,\overline{a}_m,F)$, and *(b)* parametric models $f\left[\ell_m \mid \overline{\ell}_{m-1},\overline{a}_{m-1},h\left(\psi_0\right);\phi_0\right]$, $f\left(h\left(\psi_0\right);\eta_0\right)$, and $f\left[a_m \mid \overline{\ell}_m,\overline{a}_{m-1};\alpha_0\right]$ for the above densities, where $h(\psi)$ is defined like $h(\gamma)$ except $\gamma^*\left(y,\overline{\ell}_m,\overline{a}_m,\psi\right)$ replaces $\gamma(y,\overline{\ell}_m,\overline{a}_m,F)$. It follows from (7.7) that the maximum likelihood estimates of (ϕ_0,η_0,ψ_0) are the values of $\left(\widehat{\phi},\widehat{\eta},\widehat{\psi}\right)$ that maximize

$$\prod_i \left\{\left[\frac{\partial h_i\left(\psi\right)}{\partial y_i}\right] \bullet f\left(h_i\left(\psi\right);\eta\right)\right\} \bullet \prod_{m=0}^{m=K} f\left(\ell_{m,i} \mid \overline{\ell}_{m-1,i},\overline{a}_{m-1,i},h_i\left(\psi\right);\phi\right) \qquad (7.8)$$

Since the "g"-null hypothesis (5.2) is equivalent to the hypothesis that $\psi_0 = 0$, the reparameterization (7.7) has allowed us to construct fully parametric likelihood-based tests of the "g"-null hypothesis (5.2) based on the Wald, score, or likelihood ratio test for $\psi_0 = 0$.

7.5. Estimation of $b\left(y,g\right)$

If our fully parametric likelihood-based test of the null hypothesis $\psi_0 = 0$ rejects, we would wish to employ these same parametric models to estimate $b\left(y,g\right)$ for each $g \in \mathbf{g}$. We shall accomplish this goal in two steps. First we provide a Monte Carlo algorithm which produces independent realizations of a random variable whose survivor function is $b\left(y,g\right)$.

In order to accomplish this, we shall use the function $q(h,\overline{\ell}_m,\overline{a}_m,F)$ defined recursively for all $\overline{\ell}_m$, $\overline{a}_m$ and h as follows. $q(h,\overline{\ell}_0,\overline{a}_0,F) = \gamma^{-1}\left(h,\overline{\ell}_0,\overline{a}_0,F\right)$ where the "blip-up function" $\gamma^{-1}\left(h,\overline{\ell}_k,\overline{a}_k,F\right) \equiv y$ if $\gamma\left(y,\overline{\ell}_k,\overline{a}_k,F\right) = h$. For $1 \leq k \leq m$,

$$q(h,\overline{\ell}_k,\overline{a}_k,F) = \gamma^{-1}(q(h,\overline{\ell}_{k-1},\overline{a}_{k-1},F),\overline{\ell}_k,\overline{a}_k,F).$$

Example: If $\gamma\left(y,\overline{\ell}_m,\overline{a}_m,F\right)$ is given by Eq. (7.2), $q\left(h,\overline{\ell}_k,\overline{a}_k,F\right) =$
$h - \sum\limits_{m=0}^{k}\left(2a_m + 3a_m a_{m-1} + 4a_m w_m\right)$.

MC Algorithm: Given a regime g:
Step (1): Set $v = 1$
Step (2): Draw h_v from $f_{H(\gamma)}(h)$
Step (3): Draw $\ell_{o,v}$ from $f[\ell_o \mid h_v]$
Step (4): Do for $m = 1,\ldots,K$
Step (5): Draw $\ell_{m,v}$ from $f[\ell_m \mid \overline{\ell}_{m-1,v},g(\overline{\ell}_{m-1,v}),h_v]$.
Step (6): Compute $y_{v,g} = q\left(h_v,\overline{\ell}_{K,v},g\left(\overline{\ell}_{K,v}\right),F\right)$ and return to Step (1).

Robins (1989, 1995) shows that the $(y_{1,g},\ldots,y_{v,g},\ldots)$ are independent realizations of a random variable with survivor function $b\left(y,g\right)$.

Second, since $f_{H(\gamma)}\left(h\right)$ and $f[\ell_m \mid \overline{\ell}_{m-1,v},g(\overline{\ell}_{m-1,v}),h_v]$ are unknown, in practice, we will draw h_v from $f(h;\widehat{\eta})$ and $\ell_{m,v}$ from $f[\ell_m \mid \overline{\ell}_{m-1,v},g(\overline{\ell}_{m-1,v}),h_v;\widehat{\phi}]$. Also, will we use $q^*[h_v,\overline{\ell}_{m-1,v},g(\overline{\ell}_{m-1,v}),\widehat{\psi}]$ in place of the unknown $q[h_v,\overline{\ell}_{m-1,v},g(\overline{\ell}_{m-1,v}),F]$ where $q^*[h_v,\overline{\ell}_{m-1,v},g(\overline{\ell}_{m-1,v}),\widehat{\psi}]$ is defined like $q[h_v,\overline{\ell}_{m-1,v},g(\overline{\ell}_{m-1,v}),F]$ except with $\gamma^{*-1}\left(h,\overline{\ell}_k,\overline{a}_k,\widehat{\psi}\right)$ replacing $\gamma^{-1}\left(h,\overline{\ell}_k,\overline{a}_k,F\right)$ and $\gamma^{*-1}\left(h,\overline{\ell}_k,\overline{a}_k,\widehat{\psi}\right) \equiv y$ if $\gamma^*\left(y,\overline{\ell}_k,\overline{a}_k,\widehat{\psi}\right) = h$.

Remark 7.5.1: If $\gamma\left(y, \overline{\ell}_m, \overline{a}_m, F\right) \equiv \gamma\left(y, \overline{a}_m, F\right)$ does not depend on $\overline{\ell}_m$ (i.e., there is no treatment-covariate interaction in the blip function), then, in the above algorithm for a non-dynamic regime $g = (\overline{a})$, steps (3)-(5) can be eliminated since $q\left(h, \overline{\ell}_K, \overline{a}_K, F\right) \equiv q\left(h, \overline{a}_K, F\right)$ does not depend on $\overline{\ell}_K$; as a consequence, to draw from the law of $Y_{g=(\overline{a})}$ one does not need to model the conditional density of the variables L_m. In fact, $S^{g=(\overline{a})}(y) = pr\left\{q\left[H, \overline{a}_K, F\right] > y\right\}$.

Example: If $\gamma\left(y, \overline{\ell}_m, \overline{a}_m, F\right) = y + 2a_m + 3a_m a_{m-1}$ then $q\left(h, \overline{a}_K, F\right) = h - \sum\limits_{m=0}^{K} 2a_m + 3a_m a_{m-1}$.

Remark: Eq. (7.7) is only a reparameterization. In particular, Eqs. (7.6) and (7.7) do not translate into restrictions on the joint distribution of $V = \left(Y, \overline{L}_K, \overline{A}_K\right)$ since any law for V satisfies (7.6) and (7.7). Conversely, *(i)* any function $\gamma\left(y, \overline{\ell}_m, \overline{a}_m\right)$ satisfying $\gamma\left(y, \overline{\ell}_m, \overline{a}_m\right) = 0$ if $a_m = 0$, $\partial\gamma\left(y, \overline{\ell}_m, \overline{a}_m\right)/\partial y$ is positive and continuous and *(ii)* densities $f_H(h)$, $f_{L_m}\left(\ell_m \mid \overline{\ell}_{m-1}, \overline{a}_{m-1}, h\right)$ and $f_{A_m}\left(a_m \mid \overline{\ell}_m, \overline{a}_{m-1}\right)$ together induce a unique law for $V = \left(Y, \overline{L}_K, \overline{A}_K\right)$ by (a) using $f_H(\bullet) f_{L_m}(\bullet \mid \bullet)$ and $f_{A_m}(\bullet \mid \bullet)$ to determine a joint distribution for $\left(H, \overline{L}_K, \overline{A}_K\right)$ satisfying $H \coprod A_m \mid \overline{A}_{m-1}, \overline{L}_m$ and then (b) defining Y to be $q\left(H, \overline{L}_K, \overline{A}_K\right)$ with $q(\cdot,\cdot,\cdot)$ defined in terms of $\gamma(\cdot,\cdot,\cdot)$ in the obvious manner. The joint distribution of V satisfies (7.7).

8. Semiparametric Inference in SNMs

8.1. g-Estimation of ψ_0

In this Section, we assume A_m is dichotomous. Robins (1992) discusses generalizations to multivariate A_m with (possibly) continuous components. Robins (1992) argues that one will have better prior knowledge about, and thus can more accurately model, the densities $f\left(A_m = 1 \mid \overline{L}_m, \overline{A}_{m-1}\right)$ [as in Eq. (6.4)] than the densities occurring in Eq. (7.8). Indeed, with some loss of efficiency, if A_k and L_k are discrete, we can use a saturated model in Eq. (6.4), thus eliminating all possibility of misspecification. It is for this reason we prefer to test the "g"-null hypothesis $\psi_0 = 0$ using the g-test of Sec. 6 rather than the likelihood-based test of Sec. 7.4. Here, we describe how to obtain $n^{\frac{1}{2}}$-consistent g-estimates $\widetilde{\psi}$ of ψ_0 which are based on model (6.4) and thus will be consistent with the g-test of ψ_0 in the sense that 95% confidence intervals for ψ_0 will fail to cover 0 if and only if the g-test of $\psi_0 = 0$ rejects. Specifically, we *(a)* add $\theta' Q_m(\psi)$ [rather than θQ_m] to the regressors $\alpha_0' W_m$ in (6.4) where $Q_m(\psi) = q^*\left\{H(\psi), \overline{L}_m, \overline{A}_{m-1}\right\}, q^*(\,)$ is a known vector-valued function of $\dim\psi$ chosen by the data analyst, θ is a $\dim\psi$ valued parameter; *(b)* define the G-estimate $\widetilde{\psi}$ to be the value of ψ for which the logistic regression score test of $\theta = 0$ is precisely zero; and *(c)* a 95% large sample confidence set for ψ_0 is the set of ψ for which the score test (which I call a G-test) of the hypothesis $\theta = 0$ fails to reject. (Robins, 1992). The parameter ψ is treated as a fixed constant when calculating the score test. The optimal choice of the function $q^*()$ is considered in Sec. 9.

Given $\widetilde{\psi}$, we estimate $b(y, g)$ by *(i)* finding $\widetilde{\phi}$ that maximizes (7.8) with the expression in set braces set to 1 and with ψ fixed at $\widetilde{\psi}$, and *(ii)* using the empirical distribution of $H_i\left(\widetilde{\psi}\right)$ as an estimate for the distribution of $H(\gamma)$ and *(iii)* use the MC algorithm of Sec. (7.5) to estimate $b(y, g)$ based on $\left(\widetilde{\psi}, \widetilde{\phi}\right)$ and the empirical law of $H_i\left(\widetilde{\psi}\right)$.

8.2. Subtleties in the Estimation of Direct Effect Using SNDMs

Consider again an AIDS study and suppose subjects taking aerosolized pentamidine (AP) for pneumocystis pneumonia (PCP) are less likely to take AZT treatment than other subjects because of the concern that the prophylaxis therapy may exacerbate AZT-related toxicities. In this case, the net (that is, overall) effect of AP on the outcome Y will underestimate its direct effect. In this section, we consider the estimation of the direct effect of AP using SNDMs. We redefine $A_k = (A_{Pk}, A_{Zk})$ with A_{Pk} and A_{Zk} respectively the aerosolized pentamidine and AZT dosage rates in the interval $(t_k, t_{k+1}]$. For simplicity, we will suppose that both treatments are dichotomous: $A_{Pk} = 1$ if the subject is taking prophylaxis therapy in the interval and $A_{Pk} = 0$ otherwise, with A_{Zk} similarly defined.

Let g_1, g_2, g_3, g_4 represent, respectively, the non-dynamic regimes (1) always withhold AZT and AP; (2) always withhold AZT, always take AP; (3) always take AZT, always withhold AP; and (4) always take both AZT and AP. The contrasts *(a)* $C_{12}(y) = S^{g_2}(y) - S^{g_1}(y)$ and *(b)* $C_{34}(y) = S^{g_4}(y) - S^{g_3}(y)$ represent, respectively, the direct effect AP on the distribution (survivorship function) of Y with AZT *(a)* always withheld and *(b)* always taken. We now show that even if the causal blip function $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right)$ does not depend on AP history $\overline{a}_{Pm}$, nonetheless, AP may have a direct effect on Y when AZT is always taken (i.e., $C_{34}(y) \neq 0$).

Suppose that we have a correctly specified SNDM

$$\gamma^{*}\left(y, \overline{\ell}_m, \overline{a}_m, \psi\right) = y + \psi_1 a_{Pm} + \psi_2 a_{Zm} + \psi_3 a_{Pm} w_m + \psi_4 a_{Zm} w_m, \tag{8.1}$$

say, for the causal blip function $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right)$ and w_m is a univariate function of $\overline{\ell}_m$ (such as the indicator of whether a subject has experienced a bout of PCP through t_m). Then if (3.1), (3.3), and assumption (3.10) of sequential randomization with respect to Y hold, we can estimate $S^g(y)$ for all $g \in \mathbf{g}$ and thus the contrasts $C_{12}(y)$ and $C_{34}(y)$ using the methods described in Secs. (7.5) or (8.1).

Furthermore, if $0 = \psi_{0,1} = \psi_{0,2} = \psi_{0,3} = \psi_{0,4}$, the g-null hypothesis holds by Theorem 7.1, and thus, $C_{12}(y) = C_{34}(y) = 0$ for all y. Here $\psi_{0,k}$ is the true value of the parameter ψ_k. Suppose next that $\psi_{0,1} = \psi_{0,3} = \psi_{0,4} = 0$, but $\psi_{0,2} \neq 0$. Then, by Remark 7.5.1, $C_{12}(y) = C_{34}(y) = 0$ remains true since $q(h, \overline{a}_K, F)$ depends on $\overline{a}_K$ only through $\overline{a}_{ZK}$.

More interestingly, suppose that

$$\psi_{0,1} = \psi_{0,3} = 0,\ \psi_{0,2} \neq 0,\ \psi_{0,4} \neq 0 \tag{8.2}$$

so that the blip function $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right)$ still does not depend on the AP history $\overline{a}_{Pm}$. One might wrongly suspect that if the blip function does not depend on AP history, then AP has no direct effect on survival. It is clear, however, from the MC algorithm of Sec. 7.5 that if the true values of the parameters of model (8.1) are given by (8.2), then $C_{12}(y) = 0$ for all y, but $C_{34}(y) \neq 0$ for some y, unless $f\left(\ell_m \mid \overline{\ell}_{m-1}, \overline{a}_{m-1}, h\right)$ does not depend on AP history $\overline{a}_{P(m-1)}$ for all m (i.e., $L_m \coprod \overline{A}_{P(m-1)} \mid \overline{A}_{Z(m-1)}, \overline{L}_{m-1}, H(\psi_0)$).

Remark: Heuristically, this reflects the fact that, if subjects are always on AZT, then AP may affect Y by directly affecting L_m and then L_m affects Y through the term $\psi_4 a_{Zm} w_m$ in model (8.1). We say "heuristically" because, under our assumption (3.10) of randomization with respect to Y, the "effects" of AP on L_m and of L_m on Y may not have causal interpretations.

Thus, given a SNDM that includes a non-zero A_{Zm}-$\overline{L}_m$ interaction (such as (8.1)), we can test the hypothesis $C_{12}(y) = 0$ of no direct affect of A_P on Y when A_Z is always withheld by testing whether the blip function depends on $\overline{A}_{Pm}$. However, we cannot test the hypothesis $C_{34}(y) = 0$ of no direct affect of A_P on Y when A_Z is always taken without modelling the conditional distribution of the confounders L_m given $\overline{L}_{m-1}, \overline{A}_{m-1}$ and $H(\psi_0)$.

Hence, if we were particularly interested in testing the null hypothesis $C_{34}(y) = 0$, then, in specifying our blip function, we should redefine the baseline level of AZT to be "zero" when a subject is actively taking AZT. That is, $A_{Zk} = 0$ if a subject is taking AZT in $(t_k, t_{k+1}]$ and $A_{Zk} = 1$ otherwise. If, with this redefinition, the blip function $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right)$ does not depend on AP history $\overline{a}_{Pm}$, then $C_{34}(y)$ will be 0, although, now, $C_{12}(y)$ may be non-zero.

However, if interest lies in the joint null hypothesis $C_{12}(y) = C_{34}(y) = 0$ of no direct effect of AP on Y controlling for any level of AZT treatment, use of structural nested models is inadequate since (a) $\gamma^{\dagger}\left(y, \overline{\ell}_m, \overline{a}_m\right)$ may depend on both $\overline{a}_{Pm}$ and $\overline{\ell}_m$, (b) $f\left(\ell_m \mid \overline{\ell}_{m-1}, \overline{a}_{m-1}, h\right)$ may depend on $\overline{a}_{P(m-1)}$, and yet (c) the joint null hypothesis may be true. In such a case, for essentially all non-linear SNDMs $\gamma^{*}\left(y, \overline{\ell}_m, \overline{a}_m, \psi\right)$ and parametric models $f\left(\ell_m \mid \overline{\ell}_{m-1}, \overline{a}_{m-1}, h; \phi\right)$, there would exist no parameter vector (ψ, ϕ) for which (a), (b) and (c) are true for data generated under the law governed by (ψ, ϕ). As a consequence, the true joint null hypothesis would almost certainly be rejected whatever the data. In Appendix 3, we offer a new extension of structural nested models, the direct-effect structural nested models, which overcome this difficulty by reparameterizing the joint distribution of the observables in terms of a parameter ψ_P that, when zero, implies $C_{12}(y) = C_{34}(y) = 0$.

8.3. Structural Nested Mean Models

SNDMs only apply to outcomes Y with continuous distributions. In contrast, the structural nested mean models (SNMMs) discussed in this section apply to both continuous and discrete Y. We continue to assume that $Y \equiv L_{K+1}$ is univariate. We now let $\gamma^{\dagger}\left(\overline{\ell}_m, \overline{a}_m\right) = E\left[Y_{g=(\overline{a}_m,0)} - Y_{(\overline{a}_{m-1},0)} \mid \overline{L}_m = \overline{\ell}_m, \overline{A}_m = \overline{a}_m\right]$ be the effect on the mean of Y of a final blip of treatment a_m on subjects with observed history $\left(\overline{\ell}_m, \overline{a}_m\right)$. Note $\gamma^{\dagger}\left(\overline{\ell}_m, \overline{a}_m\right) = 0$ if $a_m = 0$. Let $H_m = Y - \sum_{k=m}^{K} \gamma^{\dagger}\left(\overline{L}_k, \overline{A}_k\right)$ and redefine $H \equiv H_0$. Robins (1994) shows that, under the consistency assumption (3.1),

$$E\left[H_m \mid \overline{L}_m, \overline{A}_m\right] = E\left[Y_{g=(\overline{a}_{m-1},0)} \mid \overline{L}_m, \overline{A}_m\right] \ . \tag{8.3}$$

In particular, $E(H) = E\left[Y_{g=(0)}\right]$. Further if Eqs. (3.1), (3.3), and the sequential randomization assumption (3.10) hold, then by Theorem (3.2) and Eq. (8.3), $E\left[H_m \mid \overline{L}_m, \overline{A}_m\right] = E\left[H_m \mid \overline{L}_m, \overline{A}_{m-1}\right]$ so that H is mean independent of A_m given $\left(\overline{L}_m, \overline{A}_{m-1}\right)$, i.e.,

$$E\left[H \mid \overline{L}_m, \overline{A}_m\right] = E\left[H \mid \overline{L}_m, \overline{A}_{m-1}\right] \tag{8.4}$$

since, from its definition, H is a deterministic function of $\left(H_m, \overline{L}_m, \overline{A}_{m-1}\right)$.

We say the data follow a SNMM if $\gamma^{\dagger}\left(\overline{\ell}_m, \overline{a}_m\right) = \gamma^{*}\left(\overline{\ell}_m, \overline{a}_m, \psi_0\right)$ where $\gamma\left(\overline{\ell}_m, \overline{a}_m, \psi\right)$ is a known function depending on a finite dimensional parameter ψ satisfying $\gamma^{*}\left(\overline{\ell}_m, \overline{a}_m, \psi\right) = 0$ if $a_m = 0$ or $\psi = 0$, so $\psi_0 = 0$ represents the null hypothesis of no effect of a final blip of treatment of size a_m on the mean of Y for subjects with an observed history $\left(\overline{\ell}_m, \overline{a}_m\right)$. An example of such a function is

$$\gamma^{*}\left(\overline{\ell}_m, \overline{a}_m, \psi\right) = \psi_1 a_m + \psi_2 a_m a_{m-1} + \psi_3 a_m w_m$$

where w_m is a function of $\overline{\ell}_m$. Pseudo-SNMMs are considered in Appendix 2.

Define $H(\psi) \equiv Y - \sum_{k=0}^{K} \gamma^{*}\left(\overline{L}_k, \overline{A}_k, \psi\right)$. It follows from Eq. (8.4) and Theorem (3.2) that, given (3.1), (3.3), and the sequential randomization assumption (3.10),

$$E\left[H(\psi_0) \mid \overline{L}_m, \overline{A}_m\right] = E\left[H(\psi_0) \mid \overline{L}_m, \overline{A}_{m-1}\right] \ . \tag{8.5}$$

Robins (1994) showed that (8.5) implies that, under regularity conditions, the g-estimate $\widetilde{\psi}$ of ψ_0 of Sec. 8.1 is a CAN estimator of the parameter ψ_0 of the SNMM $\gamma\left(\overline{\ell}_m, \overline{a}_m, \psi\right)$ provided that the covariate $Q_m(\psi) = q^{*}\left\{H(\psi), \overline{L}_m, \overline{A}_{m-1}\right\}$ added to Eq. (6.4) is linear in $H(\psi)$, i.e., $q^{*}\left(H(\psi), \overline{L}_m, \overline{A}_{m-1}\right) = q_1^{*}\left(\overline{L}_m, \overline{A}_{m-1}\right) H(\psi) + q_2^{*}\left(\overline{L}_m, \overline{A}_{m-1}\right)$. It follows that, under sequential randomization, a $n^{\frac{1}{2}}$-consistent estimator of $E\left(Y_{g=(0)}\right)$ is given by $n^{-1} \sum_{i=1}^{n} H_i\left(\widetilde{\psi}\right)$. Robins (1994) shows that the g-estimate $\widetilde{\psi}$ will attain the semiparametric efficiency bound for the model when the choices of $q_1^*\left(\right)$ and $q_2^*\left(\right)$ are optimal. [However, a large sample 95 percent confidence set for ψ_0 can no longer be obtained as the set of ψ for which the score test of the hypothesis $\theta = 0$ rejects. Appropriate confidence procedures are discussed in Robins (1994).]

Further, under sequential randomization assumption (3.10), Robins (1994) shows that the mean effect of regime g on the mean of Y, $E\left[Y_g\right] - E\left[Y_{g=(0)}\right]$, is the limit as $V \to \infty$ of the following Monte Carlo procedure.

MC Procedure: For $v = (1, \ldots, V)$,
Draw for $m = (0, \ldots, K)$, $\ell_{m,v}$ recursively from $f\left[\ell_m \mid \overline{\ell}_{m-1,v}, g\left(\overline{\ell}_{m-1,v}\right)\right]$ and then compute $V^{-1}\sum_{v=1}^{V}\sum_{m=0}^{K}\gamma^\dagger\left\{\overline{\ell}_{m,v}, g\left(\overline{\ell}_{m,v}\right)\right\}$.

Thus, under sequential randomization, the g-null mean hypothesis

$$E\left(Y_g\right) = E\left(Y\right), g \in \mathbf{g} \tag{8.6}$$

holds if and only if $\gamma^\dagger\left(\overline{\ell}_m, \overline{a}_m\right) = 0$ for all m. Further, if, for each m, $\gamma^\dagger\left(\overline{\ell}_m, \overline{a}_m\right) \equiv \gamma^\dagger\left(\overline{a}_m\right)$ does not depend on $\overline{\ell}_m$, then, for a non-dynamic regime $g = \left(\overline{a}\right)$, $H = Y - \sum_{m=0}^{K}\gamma^\dagger\left(\overline{A}_m\right)$. Thus, given a SNMM $\gamma^*\left(\overline{a}_m, \psi\right)$ for $\gamma^\dagger\left(\overline{a}_m\right)$, $E\left[Y_{g=(\overline{a})} - Y_{g=(0)}\right]$ can be consistently estimated by $\sum_{m=0}^{K}\gamma^*\left(\overline{a}_m, \widetilde{\psi}\right)$.

On the other hand, if $\widetilde{\psi} \neq 0$ and either g is dynamic or $\gamma^\dagger\left(\overline{\ell}_m, \overline{a}_m\right)$ depends on $\overline{\ell}_m$, one can *(i)* specify a parametric model $f\left[\ell_m \mid \overline{\ell}_{m-1}, \overline{a}_{m-1}; \eta\right]$; *(ii)* find $\widehat{\eta}$ that maximizes $\prod_{i=1}^{n}\prod_{m=0}^{K} f\left[L_{mi} \mid \overline{L}_{(m-1)\,i}, \overline{A}_{(m-1)\,i}; \eta\right]$; *(iii)* and estimate $E\left[Y_g - Y_{g=(0)}\right]$ by
$V^{-1}\sum_{v=1}^{V}\sum_{m=0}^{K}\gamma^*\left(\overline{\ell}_{m,v}, g\left(\overline{\ell}_{m,v}\right), \widetilde{\psi}\right)$ where $\ell_{m,v}$ is drawn recursively from
$f\left[\ell_m \mid \overline{\ell}_{(m-1),v}, g\left(\overline{\ell}_{(m-1),v}\right); \widehat{\eta}\right]$. Appendix 2 provides a more complex "Monte Carlo procedure" that allows one to estimate the survivor function $S^g\left(y\right)$ under a SNMM.

We have not discussed likelihood-based methods for estimating the parameter ψ_0 of a SNMM. Rather, we have only considered g-estimation. This reflects the fact that likelihood-based inference for the parameter ψ_0 of a SNMM is somewhat complex. It is discussed in Appendix 2.

Multiplicative SNMMs: A SNMM fails to automatically impose the restriction $E\left[Y_g\right] \geq 0$ when Y_g is a non-negative random variable. However, we can use multiplicative SNMMs to impose this restriction. Specifically, redefine $\gamma^\dagger\left(\overline{\ell}_m, \overline{a}_m\right) = \ell n\left\{E\left[Y_{g=(\overline{a}_m,0)} \mid \overline{\ell}_m, \overline{a}_m\right] / E\left[Y_{g=(\overline{a}_{m-1},0)} \mid \overline{\ell}_m, \overline{a}_m\right]\right\}$. We then say the data follow a multiplicative SNMM if $\gamma^\dagger\left(\overline{\ell}_m, \overline{a}_m\right) = \gamma^*\left(\overline{\ell}_m, \overline{a}_m, \psi_0\right)$ with $\gamma^*\left(\overline{\ell}_m, \overline{a}_m, \psi\right)$ as defined above. Redefine $H = Y\exp\left\{-\sum_{m=0}^{K}\gamma^\dagger\left(\overline{L}_m, \overline{A}_m\right)\right\}$ and $H\left(\psi\right) = Y\exp\left\{-\sum_{m=0}^{K}\gamma^*\left(\overline{L}_m, \overline{A}_m, \psi\right)\right\}$. It can be proven that $E\left(H\right) = E\left(Y_{g=(0)}\right)$ and further, when (3.1), (3.3), and the sequential randomization assumption (3.10) hold, the g-null mean hypothesis (8.6) holds if and only if $\gamma^\dagger\left(\overline{\ell}_m, \overline{a}_m\right) = 0$ for all $\overline{\ell}_m, \overline{a}_m$. Furthermore, under regularity conditions the G-estimate $\widetilde{\psi}$ of ψ_0 and the estimate $n^{-1}\sum_i H_i\left(\widetilde{\psi}\right)$ of $E\left[Y_{g=(0)}\right]$ are CAN. Additionally, if, for each m, $\gamma^\dagger\left(\overline{\ell}_m, \overline{a}_m\right) \equiv \gamma^\dagger\left(\overline{a}_m\right)$ does not depend on $\overline{\ell}_m$ then $E\left[Y_{g=(\overline{a})}\right] = E\left[Y_{g=(0)}\right]\exp\left[\sum_{m=0}^{K}\gamma^\dagger\left(\overline{a}_m\right)\right]$. Thus $\ell n\left\{E\left[Y_{g=(\overline{a})}\right] / E\left[Y_{g=(0)}\right]\right\}$ can be consistently estimated by $\sum_{m=0}^{K}\gamma^*\left(\overline{a}_m, \widetilde{\psi}\right)$. However, if $\widetilde{\psi} \neq 0$ and either g is non-dynamic or $\gamma^\dagger\left(\overline{\ell}_m, \overline{a}_m\right)$ depends on $\overline{\ell}_m$, no simple analog of the above SNMM Monte Carlo procedure exists for estimating $\ell n\left[E\left(Y_g\right) / E\left(Y_{g=(0)}\right)\right]$. Rather, the more complex Monte Carlo methods described in Appendix 2 must be used.

Neither a SNMM or a multiplicative SNMM automatically imposes the true restriction $0 \leq E\left[Y_g\right] \leq 1$ when Y and Y_g are Bernoulli random variables. Unfortunately, there is no simple method to estimate the parameter ψ_0 of a “logistic” SNMM $\gamma^*\left(\overline{\ell}_m, \overline{a}_m, \psi\right)$ that imposes the restriction that $\gamma^*\left(\overline{\ell}_m, \overline{a}_m, \psi_0\right) = logit\left\{E\left[Y_{g=(\overline{a}_m,0)} \mid \overline{a}_m, \overline{\ell}_m\right]\right\} - logit\left\{E\left[Y_{g=(\overline{a}_{m-1},0)} \mid \overline{a}_m, \overline{\ell}_m\right]\right\}$.

9. g-Estimation without Sequential Randomization

If our fundamental assumption that sequential randomization (3.10) holds for $g \in \mathbf{g}$ is false, the structural blip function $\gamma^\dagger$ will differ from the pseudo-structural blip function γ. Thus if $\gamma^*\left(y, \overline{\ell}_m, \overline{a}_m, \psi\right)$ is a correctly specified SNDM, then it will not represent a correctly specified pseudo-SNDM. In this section, we assume that $\gamma^*\left(y, \overline{\ell}_m, \overline{a}_m, \psi\right)$ is a correctly specified SNDM so that $\gamma^*\left(y, \overline{\ell}_m, \overline{a}_m, \psi_0\right)$ still represents the causal effect of a final brief bit of treatment among subjects with observed history $\overline{\ell}_m, \overline{a}_m$. Now if (3.10) is false, then, according to Eq. (7.6), A_m will no longer be independent of $H\left(\psi_0\right)$ given $\overline{L}_m, \overline{A}_{m-1}$, since $H\left(\psi_0\right) = H\left(\gamma^\dagger\right)$ but $H\left(\psi_0\right) \neq H\left(\gamma\right)$. However, even when (3.10) is false, Robins (1993, App. 1) shows that the consistency assumption (3.1) implies $H\left(\gamma^\dagger\right)$ has the same distribution as $Y_{g=(0)}$. More generally, given $\left(\overline{L}_m, \overline{A}_m\right)$, $H_m\left(\gamma^\dagger\right)$ has the same conditional distribution as $Y_{g=\overline{A}_{(m-1,0)}}$. Thus it remains of interest to estimate the parameter ψ_0 of our SNDM. To do so, suppose A_m is dichotomous and we can correctly specify a logistic model

$$f\left(A_m = 1 \mid \overline{L}_m, \overline{A}_{m-1}, H\left(\psi_0\right)\right) = \left\{1 + \exp\left[-\left\{\alpha_{10}' W_m + \alpha_{20}' W_m^*\left(\psi_0\right)\right\}\right]\right\}^{-1} \tag{9.1}$$

where $W_m^*\left(\psi\right) = w_m^*\left(H\left(\psi\right), \overline{L}_m, \overline{A}_{m-1}\right)$ is a function of $H\left(\psi\right), \overline{L}_m, \overline{A}_{m-1}$ and $W_m = w_m\left(\overline{L}_m, \overline{A}_{m-1}\right)$. Note the hypothesis $\alpha_{20} = 0$ is equivalent to the hypothesis that $H\left(\psi_0\right) \coprod A_m \mid \overline{L}_m, \overline{A}_{m-1}$ and thus the hypothesis that our SNDM is also a pseudo-SNDM. The functional form of $w_m^*\left(\cdot,\cdot,\cdot\right)$ will be hard to accurately specify. Thus in practice, we suggest, as a sensitivity analysis, repeating the following analysis a number of times with different choices for $w_m^*\left(\cdot,\cdot,\cdot\right)$ and $w_m\left(\cdot,\cdot,\cdot\right)$.

Remark: Given our SNDM is correctly specified, the ability to specify model (9.1) correctly is equivalent to the ability to specify a model $f\left(A_m = 1 \mid \overline{L}_m, \overline{A}_{m-1}, Y_{g=\overline{A}_{(m-1,0)}}\right)$ since $H\left(\gamma^\dagger\right)$ is a deterministic function of $H_m\left(\gamma^\dagger\right)$ and $\overline{L}_m, \overline{A}_{m-1}$.

Given model (9.1), a G-estimate $\widetilde{\psi}$ of ψ_0 of our SNDM is obtained as in Sec. 8 except now $\theta' Q_m\left(\psi\right)$ is added to model (9.1) (with ψ_0 replaced by ψ) rather than to (6.4). $\widetilde{\psi}$ is now the value of ψ for which the score test of $\theta = 0$ is exactly 0 and a 95% confidence set for ψ_0 is still the set of ψ for which the score test of the hypothesis $\theta = 0$ fails to reject.

Under our model (9.1), the true value of ψ_0 may not be identified, the asymptotic variance of $\tilde{\psi}$ may be infinite, the score test of the hypothesis $\theta = 0$ may not reject for any value of ψ, and there may not be a unique estimate $\tilde{\psi}$. We now give a sufficient *theoretical* condition that guarantees the estimate $\tilde{\psi}$ of ψ_0 in a SNDM will indeed have infinite asymptotic variance. We then describe how this theoretical result might be used in practice.

Let $S_\psi\left(\psi, \alpha\right) = \partial \log \mathcal{L}\left(\psi, \alpha\right) / \partial \psi$ be the score for ψ for a single subject. Based on the likelihood $\mathcal{L}\left(\psi, \alpha\right) = \left\{\partial H\left(\psi\right) / \partial \psi\right\} f\left\{H\left(\psi\right)\right\} \prod_{m=0}^{K} f\left(L_m \mid \overline{L}_{m-1}, \overline{A}_{m-1}, H\left(\psi\right)\right) \prod_{m=0}^{K} f\left(A_m \mid \overline{A}_{m-1}, \overline{L}_m, H\left(\psi\right); \alpha\right)$. Note in practice, $S_\psi\left(\psi, \alpha\right)$ is a theoretical object when we have not specified models for $f\left\{H\left(\psi_0\right)\right\}$ or $f\left(L_m \mid \overline{L}_{m-1}, \overline{A}_{m-1}, H\left(\psi_0\right)\right)$. Let $Q_{opt,m}\left(\psi_0\right)$ be the unknown function $E\left[S_\psi\left(\psi_0, \alpha_0\right) \mid \overline{L}_m, \overline{A}_{m-1}, H\left(\psi_0\right), A_m = 1\right] -$
$E\left[S_\psi\left(\psi_0, \alpha_0\right) \mid \overline{L}_m, \overline{A}_{m-1}, H\left(\psi_0\right), A_m = 0\right]$. Then $\tilde{\psi}$ that uses $Q_{opt,m}\left(\psi\right)$ in place of $Q_m\left(\psi\right)$ is the most efficient g-estimator. In fact, it attains the semiparametric variance bound for the semiparametric model characterized by the SNDM $H\left(\psi_0\right)$ and the model (9.1)that leaves the law of $H\left(\psi_0\right)$ and the law of L_m

given $\left\{\overline{L}_{m-1},\overline{A}_{m-1},H\left(\psi_0\right)\right\}$ completely unspecified. Hence, if the asymptotic variance of $\tilde{\psi}$ that uses $Q_{opt,m}\left(\psi\right)$ is not finite, then no g-estimator has a finite asymptotic variance. A necessary and sufficient condition for $\tilde{\psi}$ that uses $Q_{opt,m}\left(\psi\right)$ to have infinite asymptotic variance is that, when $\psi=\psi_0$, for each m and each subject, $Q_{opt,m}\left(\psi\right)$ is a linear combination of the regressors $\left(W_m',W_m^*\left(\psi\right)'\right)'$ in model (9.1). In particular, if $W_m^*\left(\psi\right)$ equals $Q_{opt,m}\left(\psi\right)$, then all possible g-estimators $\tilde{\psi}$ will have infinite asymptotic variance. $Q_{opt,m}\left(\psi\right)$ is also optimal when doing g-estimation under model (6.4) [i.e., when it is known that $\dot{\alpha}_{20}=0$ a priori]. In that case, $\widetilde{\psi}$ that uses $Q_{opt,m}\left(\psi\right)$ will have finite asymptotic variance.

We now consider how an analyst might use this theoretical result in practice. Suppose we obtain a confidence interval based on some initial chosen function $Q_m\left(\psi\right)$. If this interval is reasonably narrow, then we have carried out a successful g-analysis. However, if our 95 percent confidence interval for ψ is too wide to be substantively useful, then either *(i)* our choice of the function Q_m was quite inefficient or *(ii)* no choice of Q_m, including the optimal choice $Q_{opt,m}$, would give usefully narrow intervals. The best approach to try to discriminate between explanations *(i)* and *(ii)* is as follows. Again specify parametric models, depending on parameter $\rho=\left(\eta',\phi'\right)'$ for the unparameterized densities in $\mathcal{L}\left(\psi,\alpha\right)$ and rewrite $\mathcal{L}\left(\psi,\alpha\right)$ as $\mathcal{L}\left(\psi,\alpha,\rho\right)$ to reflect this dependence, find the maximum likelihood estimators $\left(\widehat{\psi},\widehat{\rho},\widehat{\alpha}\right)$ based on maximizing the product over the n subjects of the $\mathcal{L}\left(\psi,\alpha,\rho\right)$, and finally construct an estimate $\widehat{Q}_{opt,m}\left(\cdot\right)$ of $Q_{opt,m}\left(\cdot\right)$ based on the distribution implied by $\left(\widehat{\psi},\widehat{\rho},\widehat{\alpha}\right)$. Now construct a new confidence interval based on g-estimation using the function $\widehat{Q}_{opt,m}\left(\psi\right)$ rather than the original $Q_m\left(\psi\right)$. The resulting estimator and 95% confidence interval is said to be a locally efficient g-estimate and interval (at the parametric submodel indexed by ρ). It is a valid confidence interval for ψ_0 even when the models parameterized by ρ are misspecified. If this 95 percent confidence interval for ψ_0 is reasonably narrow, we report this interval and conclude that option *(i)* above was true.

If the resulting confidence interval is still uselessly wide, we conclude it is likely that there is insufficient information about ψ_0 in our semiparametric model which leaves the marginal density of $H\left(\psi_0\right)$ and the conditional law of L_m given $\overline{L}_{m-1},\overline{A}_{m-1}$ and $H\left(\psi_0\right)$ unspecified. In that case we might use the maximum likelihood estimator $\widehat{\psi}$ described above. $\widehat{\psi}$ will be consistent for ψ_0 provided model (9.1) and the parametric models $f\left\{H\left(\psi_0\right);\eta\right\}$ and $f\left\{L_m\mid\overline{L}_{m-1},\overline{A}_{m-1},H\left(\psi_0\right);\phi\right\}$ are also correctly specified. Results based on $\widehat{\psi}$ should be viewed with caution because of the difficulty in correctly specifying the parametric model for L_m given $\overline{L}_{m-1},\overline{A}_{m-1},H\left(\psi_0\right)$.

Remark: The assumptions that our SNDM is correctly specified and that (9.1) identifies the parameter ψ_0 are not sufficient to identify the law of Y_g for any $g\in\mathbf{g}$ other than $g=(0)$. However, Theorem A1.2 of Robins (1993 Appendix 1) implies that the law of Y_g for any $g\in\mathbf{g}$ is identified provided our SNDM is correctly specified, Eq. (9.1) identifies the parameter ψ_0 of the SNDM, and, in addition, for $g\in\mathbf{g}$ either *(i)* $Y_g\coprod A_m\mid\overline{L}_m,\overline{A}_{m-1}=g\left(\overline{L}_{m-1}\right),H\left(\gamma^{\dagger}\right)$ [i.e., we have sequential randomization for Y_g given $\overline{L}_m$ and $H\left(\gamma^{\dagger}\right)$] or *(ii)* there is no current treatment interaction with respect to L.

Definition: We say there is no current treatment interaction with respect to L if the treatment effect transformation function $\lambda\left(y,\overline{\ell}_m,\overline{a}_m,g\right)$ does not depend on a_m for $g\in\mathbf{g}$ where, by definition, $\lambda\left(y,\overline{\ell}_m,\overline{a}_m,g\right)$ satisfies $pr\left[Y_g>\lambda\left(y,\overline{\ell}_m,\overline{a}_m,g\right)\mid\overline{\ell}_m,\overline{a}_m\right]=pr\left[Y_{g=\left(\overline{a}_{m-1},0\right)}>y\mid\overline{\ell}_m,\overline{a}_m\right]$.

Robins (1993, p. 258-260) discusses the substantive plausibility of both assumptions *(i)* and *(ii).*

10. Continuous Time SNMs

There are two difficulties with the SNDMs of Sec. 7, both of which can be solved by defining SNDMs in continuous time. First, the meaning of the parameter ψ depends on the time between measurements.

For example, the meaning of ψ depends on whether time m is a day versus a month later than time $m-1$. Thus it would be advantageous to have the parameter ψ defined in terms of the instantaneous effect of a particular treatment rate. Second, if the covariate process L and/or the treatment process A can (randomly) jump in continuous time rather than just at the pre-specified times $1, 2, \ldots, K$, the SNDMs of Sec. 7 cannot be used.

To extend SNDMs to continuous time, we assume that Y is measured at time t_{K+1}, but now a subject's covariate process $\overline{L}(t) = \{L(u); 0 \leq u \leq t\}$ and treatment process $\overline{A}(t) = \{A(u); 0 \leq u \leq t\}$ are generated by a marked point process where, for example, $A(u)$ is recorded treatment at time u. That is, *(i)* $L(t)$ and $A(t)$ have sample paths that are step functions that are right-continuous with left-hand limits; *(ii)* the $L(t)$ and $A(t)$ process do not jump simultaneously; and *(iii)* the total number of jumps K^* of the joint $\left(\overline{A}(t), \overline{L}(t)\right)$ processes in $[0, t_{K+1}]$ is random and finite, occurring at random times $T_1, \ldots, T_{K^*}$. We choose this restricted class of sample paths because their statistical properties are well understood (Arjas, 1989). Now, given treatment history $\overline{a}(t_{K+1}) = \{a(u); 0 \leq u \leq t_{K+1}\}$ on $[0, t_{K+1}]$, let $Y_{(\overline{a}(t),0)}$ be the counterfactual value of Y under the treatment history $\overline{a}^*(t_{K+1})$ where $a^*(u) = a(u)$ for $u \leq t$ and $a^*(u) = 0$ otherwise. Similarly, let $Y_{(\overline{a}(t^-),0)}$ be the counterfactual value of Y under the treatment history $\overline{a}^*(t_{K+1})$ with $a^*(u) = a(u)$ for $u < t$, $a^*(u) = 0$ otherwise. For convenience, we have suppressed the "$g =$ " in our counterfactual notation. We shall make two additional assumptions, which attempt to capture the fact that we assume that an instantaneously brief bit of treatment has a negligible effect on Y.

Assumption 1: If Y is a continuous outcome, given any history $\overline{a}(t_{K+1})$, $Y_{(\overline{a}(t),0)} = Y_{(\overline{a}(t^-),0)}$.

Assumption 2: $pr\left[Y_{\left(\overline{A}(u^-),0\right)} > y \mid \overline{L}(t), \overline{A}(t)\right]$ is continuous as a function of u.

<u>**Remark**</u>: More elegantly, we could capture the fact that a brief bit of treatment has a negligible effect on Y by the following assumption which implies Assumption 2. Given the square integrable function $a(t), t \in [0, t_{K+1}]$, let $S[a(\cdot), y, t] \equiv$
$pr\left[Y_{(\overline{a}(t_{K+1}))} > y \mid \overline{L}(t), \overline{A}(t)\right]$ so $S[\cdot,\cdot,\cdot]$ maps $L_2[0, t_{K+1}] \times R^1 \times [0, t_{K+1}]$ into R^1 where $L_2[0, t_{K+1}]$ are the set of square integrable functions on $[0, t_{K+1}]$. Our assumption is that $S[a(\cdot), y, t]$ is L_2-continuous in $a(\cdot)$. That is, for all ϵ there exists a σ such that if, for a given $a_1(\cdot)$ and $a_2(\cdot)$, the L_2 distance between $a_1(x)$ and $a_2(x)$, $\left[\int_0^{t_{K+1}} [a_1(x) - a_2(x)]^2 dx\right]^{1/2}$, is less than σ, then the absolute value of the difference between the conditional survival curves at y of $Y_{(\overline{a}_1(t_{K+1}))}$ and $Y_{(\overline{a}_2(t_{K+1}))}$, $\mid S(a_1(\cdot), y, t) - S(a_2(\cdot), y, t) \mid$, is less than ϵ. Here $\sigma = \sigma(y, t)$ may depend on (y, t).

We first shall study the simpler continuous-time structural nested mean models (SNMMs).

Continuous-Time SNMMs:

Let $V(t, h) = E\left[Y_{\left(\overline{A}(t+h^-),0\right)} - Y_{\left(\overline{A}(t^-),0\right)} \mid \overline{L}(t), \overline{A}(t)\right]$ be the mean causal effect on subjects with observed history $\left(\overline{L}(t), \overline{A}(t)\right)$ of a final blip of observed treatment
$\{A(u); t \leq u < t + h\}$ in the interval $[t, t+h)$. Note $V(t, 0) = 0$. Assumption 2 implies $V(t, h)$ is continuous in h. To be able to define the effect of an instantaneous treatment rate, we need to assume $V(t, h)$ is differentiable with respect to h.

Assumption 3: We assume *(i)* $D(t) \equiv \lim_{h \downarrow 0} V(t, h)/h$ exists for all $t \in [0, t_{K+1}]$ and *(ii)* $D(t) = \partial V(t, 0)/\partial h$ is continuous on $[T_m, T_{m+1}), m = 0, \ldots, K^* + 1$ where $T_0 \equiv 0, T_{K^*+1} \equiv t_{K+1}$.

$V(t, h)$ and $D(t)$ may be discontinuous in t at the jump times T_m because of the abrupt change in the conditioning event defining $V(t, h)$ at $t = T_m$. $D(t)\,dt$ is the effect of a last blip of observed treatment $A(t)$ sustained for "instantaneous" time dt on the mean of Y.

Define $H(t) = Y - \int_t^{t_{K+1}} D(t)\,dt$ and define H to be $H(0)$. In Appendix 0 we prove

Theorem 10.1: $E\left[H(t) \mid \overline{L}(t), \overline{A}(t)\right] = E\left[Y_{\left(\overline{A}(t^-),0\right)} \mid \overline{L}(t), \overline{A}(t)\right]$. In particular, $E(H) = E\left[Y_{(0)}\right]$.

We say the data follow a continuous-time SNMM $D(t,\psi)$ if $D(t) \equiv d\left(t, \overline{L}(t), \overline{A}(t)\right)$ equals $D(t,\psi_0) \equiv d\left(t, \overline{L}(t), \overline{A}(t), \psi_0\right)$ where ψ_0 is an unknown parameter to be estimated and $D(t,\psi)$ is a known function continuous in t on $[T_m, T_{m+1})$ satisfying $D(t,\psi) = 0$ if $\psi = 0$ or $A(t) = 0$. Define $H(t,\psi) \equiv Y - \int_t^{t_{K+1}} D(t,\psi)\, dt$. For simplicity, suppose $A(t)$ takes only the values 0 and 1 and let $\lambda\left(t \mid \overline{L}(t^-), \overline{A}(t^-)\right) dt$ be the probability that the $A(t)$ process will jump to a new state in the infinitesimal interval $[t, t+dt)$ given $\overline{A}(t^-)$ and $\overline{L}(t^-)$. We make the sequential randomization assumption that

$$\lambda\left(t \mid \overline{L}(t^-), \overline{A}(t^-), Y_{(0)}\right) = \lambda\left(t \mid \overline{L}(t^-), \overline{A}(t^-)\right). \tag{10.1}$$

Given a correctly specified Cox model, i.e.,

$$\lambda\left(t \mid \overline{L}(t^-), \overline{A}(t^-)\right) = \lambda_0(t) \exp\left[\alpha' W(t)\right] \tag{10.2}$$

where $W(t)$ is a vector function of $\left\{\overline{L}(t^-), \overline{A}(t^-)\right\}$, α is an unknown vector parameter, and $\lambda_0(t)$ is an unrestricted baseline hazard function, we obtain a G-estimate of the parameter ψ of the continuous-time SNMM $D(t,\psi)$ by adding the term $\theta' g\left(H(\psi), \overline{L}(t^-), \overline{A}(t^-)\right)$ to model (10.2) where $H(\psi) = H(0,\psi)$, $g(\cdot,\cdot)$ is a known function chosen by the investigator. Specifically, the G-estimate $\widehat{\psi}_{ge}$ is the value of ψ for which the Cox partial likelihood estimator of θ in the expanded model is zero. Then, $\widehat{\psi}_{ge}$ and $n^{-1}\sum_i H_i\left(\widehat{\psi}\right)$ will be $n^{\frac{1}{2}}$-consistent for ψ_0 and $E\left[Y_{(0)}\right]$ provided (10.1) is true, Cox model (10.2) is correctly specified, and $g\left\{H(\psi), \overline{L}(t^-), \overline{A}(t^-)\right\}$ is linear in $H(\psi)$, i.e., $g\left\{H(\psi), \overline{L}(t^-), \overline{A}(t^-)\right\} = g_1\left\{\overline{L}(t^-), \overline{A}(t^-)\right\} H(\psi) + g_2\left\{\overline{L}(t^-), \overline{A}(t^-)\right\}$. A consistent estimator of the asymptotic variance of $\widehat{\psi}_{ge}$ can be derived using methods in Robins (1994).

10.1. Continuous SNDM:

Suppose again that Y is a continuous random variable with a continuous distribution function. To describe a continuous-time SNDM, let $Q(y,t,h) \equiv q\left(y,t,h,\overline{L}(t), \overline{A}(t)\right)$ be the unique function such that $Y_{\left(\overline{A}(t+h^-),0\right)}$ and $Q\left(Y_{(A(t^-),0)}, t, h\right)$ have the same conditional distribution given $\overline{L}(t), \overline{A}(t)$. This is equivalent to

$$pr\left[Y_{\left(\overline{A}(t+h^-),0\right)} > Q(y,t,h) \mid \overline{L}(t), \overline{A}(t)\right] = pr\left[Y_{\left(\overline{A}(t^-),0\right)} > y \mid \overline{L}(t), \overline{A}(t)\right] \tag{10.3}$$

for $y \in R^1, t \in [0, t_{K+1}], h \in [0, t_{K+1} - t]$ so $Q(y,t,h)$ represents the causal effect of a final blip of the observed treatment $\{A(u); t \leq u < t+h\}$ on quantiles of Y. Note $Q(y,t,0) = y$. We now make a smoothness (differentiability) assumption.

Assumption 4: We assume that *(i)* $D(y,t) \equiv \lim_{h \downarrow 0}\{Q(y,t,h) - Q(y,t,0)\}/h$ exists and is bounded for all $(y,t) \in R^1 \times [0, t_{K+1}]$; *(ii)* further, for $(y,t) \in R^1 \times [T_m, T_{m+1}), m = 0, \ldots, K^*, D(y,t) = \partial Q(y,t,0)/\partial h$ and the matrix $\partial D(y,t)/\partial(y,t)$ has bounded and uniformly continuous entries.

Note $D(y,t)\,dt$ is the effect of a last blip of observed treatment $A(t)$ at t sustained for an instantaneous time dt on quantiles of Y. Hence $D(y,t) \equiv d\left(y, \overline{L}(t), \overline{A}(t)\right) = 0$ if $A(t) = 0$. $Q(y,t,h)$ and $D(y,t)$ may be discontinuous in t at the jump times T_m because of the abrupt change in the conditioning event defining $Q(y,t,h)$ when $t = T_m$.

Remark: It is important to note that we do not have to assume $Y_{\left(\overline{A}(t),0\right)}$ is differentiable in t for $t \neq T_m$. This is scientifically important since an unmeasured covariate process $U(t)$ may jump at t, and there may be a treatment-covariate process interaction so that the instantaneous effect on $Y_{\left(\overline{A}(t^-),0\right)}$ of a given treatment rate $A(t)$ may change at jump time t, in which case $Y_{(\overline{A}(t^-),0)}$ will not be differentiable (although still continuous). However, since $U(t)$ is unmeasured, Assumption 4 can still hold if, as we assume, the jump times for the $U(t)$ process have a continuous distribution, so that at any time the probability that the $U(t)$ process jumps at t is negligibly small.

It then follows from Theorem (2.3) of Sec. 6 of Loomis and Sternberg (1968) that *(i)* there exists a unique continuous solution $H(t) \equiv h\left(Y, t, \overline{L}(t_{K+1}), \overline{A}(t_{K+1})\right)$ to the differential equation $dH(t)/dt = D(H(t), t)$ satisfying $H(t_{K+1}) = Y$; and *(ii)* the solution is a continuous function of (Y, t) on $R^1 \times [0, t_{K+1}]$ and $\partial h\left(t, Y, \overline{L}(t_{K+1}), \overline{A}(t_{K+1})\right)/\partial(Y, t)$ exists and is bounded and continuous on $R^1 \times [T_m, T_{m+1})$. Hence the Jacobian $\partial H(0)/\partial Y$ for the transformation from Y to $H(0)$ exists with probability 1. The strong smoothness conditions in Assumption 4*(ii)* were required to guarantee the existence of this Jacobian. Our main result is Theorem (8.2).

<u>Theorem</u> <u>10.2</u>: $H(t)$ and $Y_{(\overline{A}(t^-),0)}$ have the same conditional distribution given $\left(\overline{L}(t), \overline{A}(t)\right)$. In particular, $H \equiv H(0)$ and $Y_{(0)}$ have the same marginal distribution.

Theorem (10.2) is proved in Appendix 0 for the special case in which there is local rank preservation, i.e., $Q\left(Y_{(A(t^-),0)}, t, h\right) = Y_{(\overline{A}(t+h^-),0)}$ with probability one. Although I am nearly certain that Theorem (10.2) holds without local rank preservation by a coupling argument, my current proof attempt still suffers from unresolved technical problems.

We say the data follows a continuous-time SNDM $D(y, t, \psi)$ if $D(y, t, \psi) = D(y, t, \psi_0)$ where ψ_0 is an unknown parameter and $D(y, t, \psi) \equiv d\left(y, t, \overline{L}(t), \overline{A}(t), \psi\right)$ is a known function satisfying *(i)* $D(y, t, 0) = 0$, *(ii)* $D(y, t, \psi) = 0$ if $A(t) = 0$, and *(iii)* Assumption 4*(ii)* holds for each fixed value of ψ. It then follows if the explainable non-random non-compliance assumption (10.1) holds, the Cox model (10.2), and the continuous-time SNDM model $D(y, t, \psi)$ are correctly specified, then the estimators $\widehat{\psi}_{ge}$ and $n^{-1}\sum_i H_i\left(\widehat{\psi}_{ge}\right)$ will be $n^{\frac{1}{2}}$-consistent for ψ_0 and $E\left[Y_{(0)}\right]$ respectively, even when the function $g\left(H(\psi), \overline{L}(t^-), \overline{A}(t^-)\right)$ to be added to the Cox model (10.2) is not linear in $H(\psi)$.

11. Faithfulness: From Association to Causation by Philosophy?

Except when there has been physical randomization (e.g., a sequential randomized trial was performed), the methods described in this paper do not allow one to claim that associations found in the data are causal without prior untestable assumptions, such as the assumption (3.10) of no unmeasured confounders. Assumption (3.10) will never be exactly true. Whether it may nonetheless serve as a reasonable working hypothesis will depend critically both on the covariates $\overline{L}_K$ available for data analysis and on the substantive issue under investigation. It is this latter reason that epidemiologic studies need to be conducted by subject matter experts.

In stark contrast, Pearl and Verma (PV) (1991) and SGS (1993) argue that one can go from "association" to "causation" without any subject matter knowledge based on a "philosophical" assumption, the assumption of faithfulness or stability. In fact, SGS have written a computer program, Tetrad, that searches epidemiologic data bases for causal relations by applying the faithfulness assumption. I will show that their argument is unconvincing. I will argue that in observational epidemiologic studies, one cannot, even in principle, go from association to causation without strong subject matter knowledge.

To demonstrate why their argument fails, I will consider the following simple example. The causal DAG in Figure 5 is a non-parametric recursive structural equation (SEM) model (Pearl, 1995). The arrows indicate the potential causal associations between variables A, B, C, V, and W. Variables V and W are unobserved, variables A, B, and C are observed. As indicated on the graph, we assume that we know the temporal ordering: A then B then C. Our goal is to determine whether B causes C. To fix ideas, one might think of C as lung cancer, B as alcohol consumption, A as city of birth, V as race, and W as cigarette consumption. We wish to know whether alcohol causes lung cancer.

A linear recursive SEM is a special case of a non-parametric recursive SEM in which all disturbances (error terms) are normal with mean zero. For this special case, the lower case letters in Figure 5 denote path coefficients. If, as a fact of nature, a path coefficient is zero, then the corresponding arrow is "missing." For example, if B causes C, the path coefficient b differs from zero. In contrast, if B does not cause C, b equals zero and the arrow from B to C can be removed.

Following SGS, we first assume that we have essentially an unlimited amount of data on variables A, B, and C (i.e., we have an infinitely large study population). Suppose in the data all two-way correlations are non-zero but A and C are independent given B, i.e., $A \coprod C \mid B$. Given these data associations, PV and SGS would invoke the philosophical faithfulness assumption and conclude that B causes C regardless of the real world variables represented by A, B, and C. That is, they would go from association to causation without subject matter knowledge and they would do so even if the magnitude of the correlation between B and C was very small, say $\rho = 10^{-5}$.

PV and SGS argue as follows. For simplicity, without loss of generality, we will restrict attention to the special case of the linear recursive SEM. When B does not cause C (i.e., $b = 0$), PV and SGS note that the only way that we could have $A \coprod C \mid B$ and $B \not\coprod C$ is:

Explanation (1): Both W and V are confounders ($w_1 w_2 \neq 0$ and $v_1 v_2 \neq 0$) and $b = 0$, but the magnitudes of the path coefficients are perfectly balanced in such a way to make A and C independent given B.

On the other hand, if B does cause C (i.e., $b \neq 0$), PV and SGS note that we would have $A \coprod C \mid B$ and $B \not\coprod C$ if the following explanation held.

Explanation (2): Both W and V are non-confounders ($w_1 w_2 = v_1 v_2 = 0$) and $b \neq 0$.

SGS and PV then proceed to rule out explanation (1) (which is always a logical mathematical possibility) by the following "faithfulness" argument. They argue that, since the subset $\{(w_1, w_2, b_1, b_2)\}$ of path coefficients that will make $A \coprod C \mid B$ when $w_1 w_2 \neq 0$ and $v_1 v_2 \neq 0$ has Lebesgue measure zero in R^4, the "prior" probability of explanation (1) occurring is zero (since explanation (1) requires such "fortuitous" values of the path coefficients). I fully agree with SGS and PV that the prior probability of Explanation (1) is zero.

However, SGS and PV then conclude that explanation (2) must be the proper explanation and thus conclude B causes C. It is this last step with which I disagree. This step can only be justified if the prior probability that V and W are both non-confounders ($v_1 v_2 = 0$ and $w_1 w_2 = 0$) is greater than zero. Otherwise, both explanations (1) and (2) are events of probability zero and so their relative likelihood is not defined without further assumptions. Now the Lebesgue measure of the event "$v_1 v_2 = 0$ and $w_1 w_2 = 0$" is also zero, since v_1, v_2,w_1, and w_2 take values in the continuum $[-1, 1]$. It follows that PV and SGS are making the "hidden" assumption that the event that a path coefficient has value zero (i.e., that an arrow is missing on the graph) has a non-zero probability. On this "hidden" assumption, explanation (2) does have a positive probability, and we can conclude that B causes C. Now I do agree with PV and SGS that, for any two given variables, the probability that the path coefficient between them is zero (i.e., neither variable causes the other) is non-zero. But in the epidemiologic example in which we are trying to determine whether drinking causes lung cancer, the unmeasured variables W and V in Fig. 5 do not represent single variables but rather represent all possible unmeasured common causes of B and C and A and C respectively. If there should exist any common cause of B and C or of A and

C, explanation (2) is false.

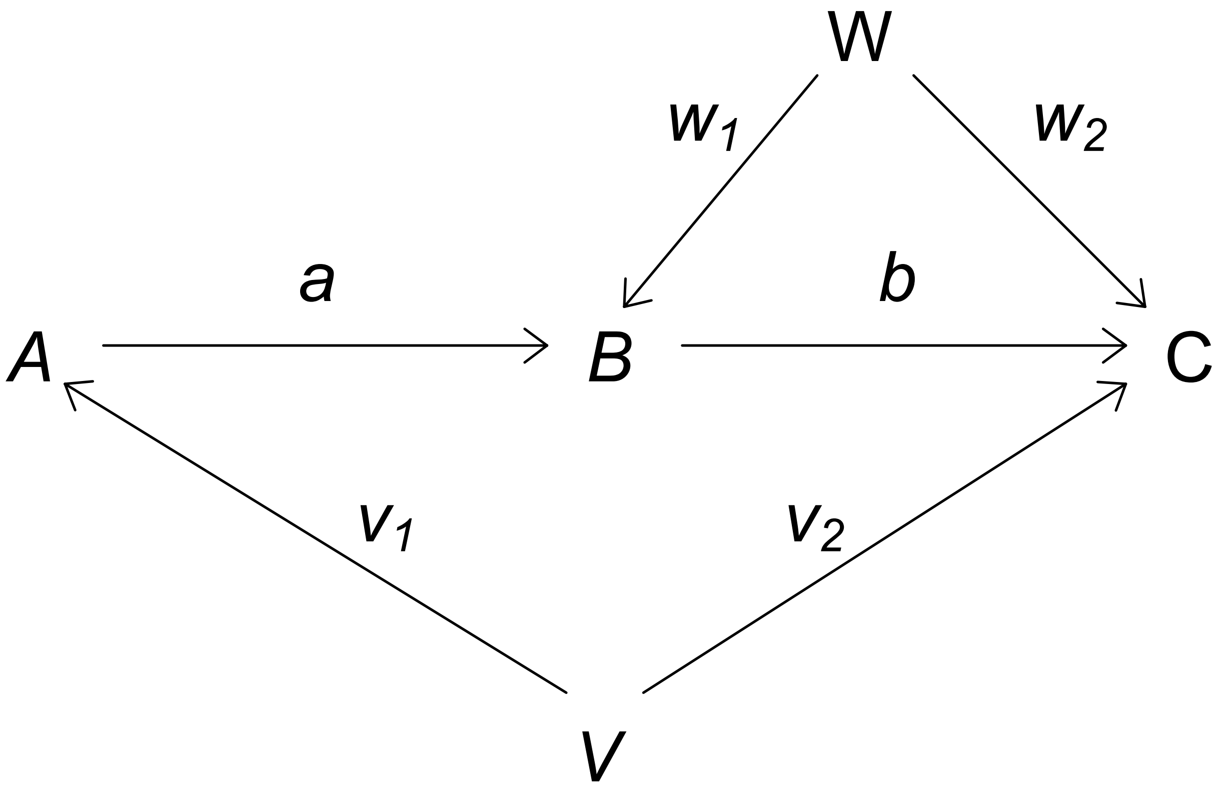


Figure 5: A Causal DAG representing a linear structural equation model with normal errors. Small letters are path coefficients. Variables *W* and *V* are unobserved. Variables *A*, *B*, and *C* are observed.

Now I (and every other epidemiologist I know) believe that, given a sufficiently large sample size, a test for independence between any two variables recorded in an observational study will reject due to the existence of unmeasured common causes (confounders). Thus, we teach epidemiology students that, with certainty (i.e., with prior probability 1), explanation (2) will not occur (once we understand that W and V represent the conglomeration of all unmeasured common causes). It follows that we cannot conclude explanation (2) is to be preferred to explanation (1). Thus we cannot conclude that B causes C.

Suppose, however, that, although Explanations (1) and (2) both have zero prior probability, we believe under further reasonable assumptions and with respect to some measure, the relative likelihood of Explanation (1) is infinitely greater than that of Explanation (2). We now argue that even so, nonetheless, in realistic studies with their finite sample size, we should still not invoke the "faithfulness assumption" and conclude that B causes C when, in the data, $A \coprod C \mid B$.

Consider now a realistic study, even a very large one, with its finite sample size. Then, even if, in the data, A is exactly independent of C given B, nonetheless, due to sampling variability, there will exist a confidence interval around the null estimate of the conditional association parameter between A and C given B. Since, as argued above, our prior probability that in truth A and C are independent given B is zero [explanations (1) and (2) both have prior probability zero], our *posterior* odds that $A \coprod C \mid B$ is exactly zero will be zero, compared to the hypothesis that A and C are dependent given B with their conditional association parameter lying in the small confidence interval surrounding zero. Thus, again, we would not necessarily conclude that B causes C since the event that we "almost" have balanced confounding and $b = 0$ does not have a prior probability of 0 and thus its posterior probability need not be small. However, if the BC correlation was large and, based on subject matter knowledge, we believed that the magnitude of the confounding by the conglomerate variables W and V was small, then we would conclude, based on a formal Bayesian analysis, that b was likely non-zero. This analysis would, of course reflect our subject matter knowledge and would not be based on any subject-matter-independent philosophical principle.

To conclude, I would recommend against using the computer program Tetrad as a tool for searching epidemiologic data bases for causal associations, since its search strategy is based solely on the implications of the faithfulness assumption. Finally, I wish to note that my argument against the use of the faithfulness assumption in analyzing epidemiologic data is not an argument against the faithfulness assumption in analyzing data in simple, stereotyped environments where the number of potential unmeasured confounders conglomerated in W and V is known to be small. Thus, in AI programs designed to allow robots to learn from data obtained in a rather stereotyped environment, the faithfulness assumption and thus a Tetrad-like program might be useful.

Appendix 0: Continuous Time SNMs

Proof of Theorem 10.1: Let $M(u,t) = \lim_{\Delta t \downarrow 0} \{Z(u+\Delta t, t) - Z(u,t)\}/\Delta t$ where $Z(u,t) \equiv E\left[Y_{\left(\overline{A}(u^-),0\right)} \mid \overline{L}(t), \overline{A}(t)\right]$. Note $M(u,t) = E\left[D(u) \mid \overline{L}(t), \overline{A}(t)\right]$ and thus, by Assumption 3, is continuous in t except at $T_m, m = 1, \ldots, K^*$. Hence $E\left[H(t) \mid \overline{L}(t), \overline{A}(t)\right] = E\left[Y \mid \overline{L}(t), \overline{A}(t)\right] - E\left[\int_t^{t_{K+1}} D(u)\,du \mid \overline{L}(t), \overline{A}(t)\right] = E\left[Y \mid \overline{L}(t), \overline{A}(t)\right] - \left\{\int_t^{t_{K+1}} M(u,t)\,du\right\} = E\left[Y \mid \overline{L}(t), \overline{A}(t)\right] -$

$$\left\{E\left[Y \mid \overline{L}(t), \overline{A}(t)\right] + \sum_{\{m; T_m > t\}} \{Z(T_m^-, t) - Z(T_m, t)\} - Z(t,t)\right\} = Z(t,t) = E\left[Y\left(\overline{A}(t^-), 0\right) \mid \overline{L}(t), \overline{A}(t)\right]$$

where the last equality is definitional, the second to last is by the fact that, by Assumption 2, $Z(u,t)$ is continuous in u, so $Z(T_m^-, t) = Z(T_m, t)$, the third to last is by the facts that $M(u,t) = \partial Z(u,t)/\partial u$ for $u \in [T_m, T_{m+1})$ and $Z(t_{K+1}, t) = E\left[Y \mid \overline{L}(t), \overline{A}(t)\right]$ by the consistency assumption.

Proof of Theorem 10.2 under Local Rank Preservation: By Theorem (2.3) of Chpt. 6 of Loomis and Sternberg (1968), there is a unique continuous solution to $dH(t)/dt = D(H(t), t), t \in [T_m, T_{m-1})$ satisfying $H(t_{K+1}) = Y$. We now show that under local rank preservation $Y_{\left(\overline{A}(t^-),0\right)}$ satisfies these conditions as a function of t. First by Assumption (4) and local rank preservation *(i)* $d\left(Y_{\left(\overline{A}(t^-),0\right)}\right)/dt$ exists and equals $D\left(Y_{\left(\overline{A}(t^-),0\right)}, t\right)$ on $[T_m, T_{m+1})$, *(ii)* $Y_{(A(t^-),0)}$ is continuous in t by Assumption (1) and (2), and *(iii)* $Y_{\left(\overline{A}(t^-),0\right)}$ evaluated at $t = t_{K+1}$ is Y by the consistency assumption. Note that, given Assumption (4), the additional assumption of (local) rank preservation implies that $Y_{\left(\overline{A}(t^-),0\right)}$ is continuously differentiable for $t \in [T_m, T_{m+1})$. [Note the assumption of local rank preservation is incompatible with the existence of the unmeasured covariate process $U(t)$ that interacts with treatment described in the Remark of Sec. 10.1.]

Appendix 1:

Proof of Theorem 3.2:

It will be sufficient to prove the rightmost equality in (3.8). In the following, set $\overline{A}_{m-1}$ to be $g\left(\overline{L}_{m-1}\right)$. If $f\left(\overline{L}_m, \overline{A}_{m-1}\right) \neq 0$, define $b^*\left(y, \overline{L}_m, \overline{A}_{m-1}\right) = E\left[I\left(Y_g > y\right) \mid \overline{L}_m, \overline{A}_{m-1}\right]$ where the dependence on g has been suppressed. Now if $f\left(\overline{L}_m, \overline{A}_{m-1}, A_m = g\left(t_m, \overline{L}_m\right)\right) \neq 0$, then $b^*\left(y, \overline{L}_m, \overline{A}_{m-1}\right) =$
$E\left[b^*\left(y, \overline{L}_{m+1}, \overline{A}_m\right) \mid \overline{L}_m, \overline{A}_{m-1}, A_m = g\left(t_m, \overline{L}_m\right)\right] \equiv$
$\int b^*\left(y, \ell_{m+1}, \overline{L}_m, A_m = g\left(t_m, \overline{L}_m\right), \overline{A}_{m-1}\right) f\left[\ell_{m+1} \mid \overline{L}_m, A_m = g\left(t_m, \overline{L}_m\right), \overline{A}_{m-1}\right] d\mu\left(\ell_{m+1}\right)$
since, by the sequential randomization assumption (3.10),
$b^*\left(y, \overline{L}_m, \overline{A}_{m-1}\right) \equiv E\left[I\left(Y_g > y\right) \mid \overline{L}_m, \overline{A}_m\right] = E\left[I\left(Y_g > y\right) \mid \overline{L}_m, \overline{A}_{m-1}, A_m = g\left(t_m, \overline{L}_m\right)\right]$
$= E\left\{E\left[I\left(Y_g > y\right) \mid \overline{L}_{m+1}, \overline{A}_m\right] \mid \overline{L}_m, \overline{A}_{m-1}, A_m = g\left(t_m, \overline{L}_m\right)\right\}$ by iterated expectations. Now putting $\overline{A}_{k-1} = g\left(\overline{L}_{k-1}\right)$ invoking (3.3) and arguing recursively, we then obtain
$b^*\left[y, \overline{L}_k, g\left(\overline{L}_{k-1}\right)\right] = \int \cdots \int b^*\left(y, \overline{L}_K, g\left(\overline{L}_{K-1}\right)\right) \prod\limits_{m=k+1}^{K} f\left[\ell_m \mid \overline{\ell}_{m-1}, g\left(\overline{\ell}_{m-1}\right)\right] d\mu\left(\ell_m\right)$. But, by (3.10) and (3.3), $b^*\left(y, \overline{\ell}_K, g\left(\overline{\ell}_{K-1}\right)\right) = E\left[I\left(Y_g > y\right) \mid \overline{\ell}_K, g\left(\overline{\ell}_K\right)\right] = E\left[I\left(Y > y\right) \mid \overline{\ell}_K, g\left(\overline{\ell}_K\right)\right] \equiv S\left(y \mid \overline{\ell}_K, g\left(\overline{\ell}_K\right)\right)$ by the consistency assumption (3.1).

Appendix 2:

Further Results on SNMMs and Pseudo-SNMMs

The semiparametric model induced by (3.1), (3.3), (3.10) and the SNMM $\gamma^*\left(\overline{\ell}_m, \overline{a}_m, \psi\right)$ restricts the joint distribution of the observables $\overline{V} = \left(\overline{L}, \overline{A}\right)$ only through the restriction (8.5). We now construct a pseudo-SNMM that obeys this restriction without reference to counterfactuals. Let $b\left(g\right) = \int y dF\left(y, g\right)$ and $b\left(\overline{\ell}_m, g\right) = \int y \; dF\left(y, \overline{\ell}_m, g\right)$ where $dF\left(y, g\right)$ and $dF\left(y, \overline{\ell}_m, g\right)$ are the measures on Y induced by $b\left(y, g\right)$ and $b\left(y, \overline{\ell}_m, g\right)$. For example, if Y has differentiable conditional distribution functions, then $dF\left(y, g\right) = \left\{-\partial b\left(y, g\right) / \partial y\right\} \; dy$ and $dF\left(y, \overline{\ell}_m, g\right) = -\left\{\partial b\left(y, \overline{\ell}_m, g\right) / \partial y\right\} dy$. Define $\gamma\left(\overline{\ell}_m, \overline{a}_m, F\right) = b\left[\overline{\ell}_m, g = \left(\overline{a}_m, 0\right)\right] - b\left[\overline{\ell}_m, g = \left(\overline{a}_{m-1}, 0\right)\right]$ where F denotes the dependence on the joint distribution F of the observables V. Set $H\left(\gamma\right) = Y - \sum\limits_{m=0}^{K} \gamma\left(\overline{\ell}_m, \overline{a}_m, F\right)$. Then it is straightforward to prove both

$$E\left[H\left(\gamma\right) \mid \overline{L}_m, \overline{A}_m\right] = E\left[H\left(\gamma\right) \mid \overline{L}_m, \overline{A}_{m-1}\right] \tag{A2.1}$$

and the following "g"-null mean theorem.
"g"-null mean theorem: $b\left(g_1\right) = b\left(g_2\right)$ for all $g_1, g_2 \in \mathbf{g}$ if and only if $\gamma\left(\overline{\ell}_m, \overline{a}_m, F\right)$ =0 for all $\left(\overline{\ell}_m, \overline{a}_m\right)$.

We say the data follow a pseudo-SNMM $\gamma\left(\overline{\ell}_m, \overline{a}_m, \psi\right)$ if the definition of a SNMM is satisfied with $\gamma\left(\overline{\ell}_m, \overline{a}_m, F\right)$ replacing $\gamma^{\dagger}\left(\overline{\ell}_m, \overline{a}_m\right)$. It follows from Theorem (3.2) that when (3.1), (3.3), and (3.10) are true, then $\gamma^*\left(\overline{\ell}_m, \overline{a}_m, \psi\right)$ is a pseudo-SNMM if and only if it is a SNMM.

Our next goal is to reparameterize the likelihood function of the observables V in such a way that the restriction (A2.1) is naturally incorporated. Note that simply writing $f_V\left(y, \overline{\ell}_K, \overline{a}_K\right) =$
$f_{H(\gamma)|\overline{L}_K, \overline{A}_K}\left(h \mid \overline{\ell}_K, \overline{a}_K\right) f\left(\overline{\ell}_K, \overline{a}_K\right)$ fails to impose the restriction (A2.1). To impose this restriction, for notational convenience, define $\overline{X}_m = \left(\overline{L}_m, \overline{A}_m\right)$. Consider the following algebraic decomposition of $H\left(\gamma\right)$.

$$\begin{gathered} H\left(\gamma\right) = \left\{H\left(\gamma\right) - E\left[H\left(\gamma\right) \mid \overline{X}_K\right]\right\} + \\ \left\{\sum_{m=0}^{K} E\left(H\left(\gamma\right) \mid \overline{X}_m\right) - E\left(H\left(\gamma\right) \mid \overline{X}_{m-1}\right)\right\} + E\left(H\left(\gamma\right)\right) \equiv \\ \sigma + \sum_{m=0}^{K} \nu\left(\overline{X}_m\right) + \beta_0 \end{gathered} \tag{A2.2}$$

with $\sigma \equiv H(\gamma) - E\left(H(\gamma) \mid \overline{X}_K\right), \nu\left(\overline{X}_m\right) = E\left[H(\gamma) \mid \overline{X}_m\right] - E\left[H(\gamma) \mid \overline{X}_{m-1}\right], \beta_0 = E\left(H(\gamma)\right)$. The restriction (A2.1) implies that $\nu\left(\overline{X}_m\right) \equiv \nu\left(L_m, \overline{X}_{m-1}\right)$ does not depend on A_m. Thus, we can write,

$$\sigma \equiv H(\gamma) - \sum_{m=0}^{K} \nu\left(L_m, \overline{X}_{m-1}\right) - \beta_0 \tag{A2.3}$$

with

$$E\left[\nu\left(L_m, \overline{X}_{m-1}\right) \mid \overline{X}_{m-1}\right] = 0 \tag{A2.4}$$

and

$$E\left[\sigma \mid \overline{X}_K\right] = 0\ . \tag{A2.5}$$

Conversely, given a distribution function for the random variables $\overline{X}_K \equiv \left(\overline{A}_K, \overline{L}_K\right)$, any constant β_0, any random variable σ satisfying (A2.5), and any function $\nu(\cdot, \cdot)$ satisfying (A2.4), then

$$H \equiv \sigma + \sum_{m=0}^{K} \nu\left(L_m, \overline{X}_{m-1}\right) + \beta_0 \tag{A2.6}$$

is mean independent of A_m given $\left(\overline{L}_m, \overline{A}_{m-1}\right)$. Hence, given a function $\gamma\left(\overline{\ell}_m, \overline{a}_m\right)$ satisfying $\gamma\left(\overline{\ell}_m, \overline{a}_m\right) = 0$ when $a_m = 0$, if we set

$$Y = H + \sum_{m=0}^{K} \gamma\left(\overline{L}_m, \overline{A}_m\right)$$

with H defined by (A2.6), then the law of $V = \left(Y, \overline{L}_K, \overline{A}_K\right)$ will have a joint distribution with $\gamma\left(\overline{\ell}_m, \overline{a}_m, F\right)$ equal to the chosen function $\gamma\left(\overline{\ell}_m, \overline{a}_m\right)$. Hence the data follow a pseudo-SNMM $\gamma^*\left(\overline{\ell}_m, \overline{a}_m, \psi\right)$ [i.e., $\gamma^*\left(\overline{\ell}_m, \overline{a}_m, \psi_0\right) = \gamma\left(\overline{\ell}_m, \overline{a}_m, F\right)$] if and only if there exists a value ψ_0 of the parameter ψ, a constant β_0, a random variable σ, and a function $\nu(\cdot, \cdot)$ [that satisfy (A2.5) and (A2.4) respectively] such that $H(\psi_0) \equiv Y - \sum\limits_{m=0}^{K} \gamma^*\left(\overline{\ell}_m, \overline{a}_m, \psi_0\right)$ is equal to the RHS of (A2.6).

Likelihood Function: Since (A2.4) implies there exists a unique function $\nu^*\left(L_m, \overline{X}_{m-1}\right)$ such that $\nu\left(L_m, \overline{X}_{m-1}\right) = \nu^*\left(L_m, \overline{X}_{m-1}\right) - E\left[\nu^*\left(L_m, \overline{X}_{m-1} \mid \overline{X}_{m-1}\right)\right]$ and $\nu^*\left(0, \overline{X}_{m-1}\right) = 0$, the subject-specific likelihood function of a pseudo-SNMM (and thus for a SNMM under sequential randomization) can be written

$$\mathcal{L}(\psi, \eta) \equiv f_V\left(Y, \overline{X}_K; \psi, \eta\right) =$$
$$\left\{\frac{\partial \sigma(\psi_0, \eta_0)}{\partial H(\psi_0)} \frac{\partial H(\psi_0)}{\partial Y}\right\} f\left(\sigma(\psi, \eta) \mid \overline{X}_K; \eta_1\right) \prod_{m=0}^{K} f\left(A_m \mid L_m, \overline{X}_{m-1}; \eta_4\right) f\left(L_m \mid \overline{X}_{m-1}; \eta_3\right),$$

with

$$\sigma(\psi, \eta) \equiv H(\psi) - \eta_5 - \sum_{m=0}^{K} \left\{\nu^*\left(L_m, \overline{X}_{m-1}; \eta_2\right) - \int \nu^*\left(\ell_m, \overline{X}_{m-1}; \eta_2\right) dF\left(\ell_m \mid \overline{X}_{m-1}; \eta_3\right)\right\} \tag{A2.7}$$

subject to the restrictions that

$$\int dF\left(t \mid \overline{X}_K; \eta_1\right) = 0 \tag{A2.8}$$

and

$$\nu^*\left(0,\overline{X}_{m-1};\eta_2\right)=0 \tag{A2.9}$$

Here $\eta=(\eta_1,\ldots,\eta_5)'$ are unknown parameters; η_5 is an unknown constant (that plays the role that β_0 did in A2.2); η_4 indexes all conditional laws of A_m given $\left(L_m,\overline{X}_{m-1}\right)$; η_3 indexes all possible laws of L_m given $\overline{X}_{m-1}$; η_1 indexes all laws of $\sigma\equiv\sigma(\psi_0,\eta_0)$ given $\overline{X}_K$ such that $E\left[\sigma\mid\overline{X}_K\right]=0$ (which implies (A2.8)); and η_2 indexes all possible functions $\nu^*(\cdot,\cdot;\eta_2)$ of $\left(L_m,\overline{X}_{m-1}\right)$ satisfying (A2.9). The Jacobian $\left\{\dfrac{\partial\sigma(\psi_0,\eta_0)}{\partial H(\psi_0)}\dfrac{\partial H(\psi_0)}{\partial Y}\right\}$ for the transformation from Y to σ is equal to 1.

It follows that fully parametric likelihood-based inference for the parameter ψ of a pseudo-SNMM is performed by *(i)* choosing parametric models for
$f(\cdot\mid\cdot;\eta_1)$, $f\left(\overline{L}_m\mid\overline{X}_{m-1};\eta_3\right)$ and for $\nu^*(\cdot,\cdot;\eta_2)$ indexed by finite dimensional parameters ω_1, ω_3, and ω_2, *(ii)* maximizing $\prod_i\mathcal{L}_i(\psi,\omega,\eta_4,\eta_5)$ with respect to (ψ,ω,η_5) to obtain $\left(\widehat{\psi},\widehat{\omega},\widehat{\eta}_5\right)$. [Note that the MLE $\left(\widehat{\psi},\widehat{\omega},\widehat{\eta}_5\right)$ does not depend on whether $f\left(A_m\mid\overline{L}_m,\overline{X}_{m-1};\eta_4\right)$ is completely unknown or follows a parametric submodel such as $f\left(A_m\mid\overline{L}_m,\overline{X}_{m-1};\alpha\right)$ of Eq. (6.4).] An alternative to fully parametric likelihood-based inference is to specify a model [such as (6.4)] for $f\left(A_m\mid\overline{L}_m,\overline{X}_{m-1}\right)$ and obtain a g-estimate $\widetilde{\psi}$ of ψ_0. Then one can obtain estimates of $(\widetilde{\omega},\widetilde{\eta}_5)$ by maximizing $\prod_i\mathcal{L}_i\left(\widetilde{\psi},\omega,\eta_4,\eta_5\right)$ with respect to ω and η_5.

Drawing From the Law of Y_g Under a SNMM $\gamma^*\left(\overline{\ell}_m,\overline{a}_m,\psi\right)$:

Given the true values (ψ_0,η_0) of (ψ,η) to obtain V-independent draws $y_{g,v}$ from the distribution $b(y,g)$ [and, thus, under (3.1), (3.3) and (3.10) from the law of Y_g], we proceed as follows.
Monte Carlo Procedure:
For $v=(1,\ldots,V)$
1). Recursively for $m=(0,\ldots,K)$, draw ℓ_{mv} from $f\left(\ell_m\mid\overline{\ell}_{(m-1)v},g\left(\overline{\ell}_{(m-1)v}\right);\eta_{0,3}\right)$;
2). Draw σ_v from $f\left(\sigma\mid\overline{\ell}_{Kv},g\left(\overline{\ell}_{Kv}\right);\eta_{0,1}\right)$;
3). Compute $h_v=\sigma_v+\eta_{0,5}+\sum\limits_{m=0}^{K}\{\nu^*\left(\ell_{mv},\overline{\ell}_{(m-1)v},g\left(\overline{\ell}_{(m-1)v}\right);\eta_{0,2}\right)-$
$\int\nu^*\left(\ell_m,\overline{\ell}_{(m-1)v},g\left(\overline{\ell}_{(m-1)v}\right);\eta_{0,2}\right)f\left(\ell_m\mid\overline{\ell}_{(m-1)v},g\left(\overline{\ell}_{(m-1)v}\right);\eta_{0,3}\right)d\mu(\ell_m)\}$;
4). Compute $y_{g,v}=h_v+\sum\limits_{m=0}^{K}\gamma\left(\overline{\ell}_{mv},g\left(\overline{\ell}_{mv}\right),\psi_0\right)$.

In practice, we replace (ψ_0,η_0) with suitable estimates.

Multiplicative SNMMs:

Results of this Appendix go over straightforwardly to multiplicative SNMMs. In particular, when (3.1), (3.3), and the sequential randomization assumption (3.10) hold, the likelihood function for a multiplicative SNMM is as above. However, the Jacobian $\{\partial\sigma(\psi_0,\eta_0)/\partial H(\psi_0)\}\{\partial H(\psi_0/\partial Y)\}$ now equals $\exp\left\{-\sum_{m=0}^{K}\gamma^*\left(\overline{L}_m,\overline{A}_m,\psi_0\right)\right\}$. Furthermore, the Monte Carlo procedure to draw from the law of Y_g is as above, except Step 4 is replaced by the following:
4′). Compute $y_{g,v}=h_v\exp\left\{\sum_{m=0}^{K}\gamma\left(\overline{\ell}_{mv},g\left(\overline{\ell}_{mv}\right),\psi_0\right)\right\}$.

Appendix 3

Suppose, as in Sec. 8.2, $Y\equiv L_{K+1}$ and $A_m=(A_{Pm},A_{Zm})$ is comprised of two distinct treatments, A_{Pm} and A_{Zm}. In this Appendix, we provide a new extension of SNMs, the direct-effect SNMs, for which, under sequential randomization (3.10), there exists a parameter ψ_P that takes the value zero if and only if the *direct-effect g-null hypothesis* of no direct effect of $\overline{a}_P$ on Y controlling for any $\overline{a}_Z$ treatment holds. To formalize this null hypothesis, let

$$\mathbf{g}_P=\left\{g\in\mathbf{g};g\left(t_k,\overline{\ell}_k\right)\equiv\left(g_P\left(t_k,\overline{\ell}_k\right),a_{Zk}\right)\ \textit{has the same component}\ a_{Zk}\ \text{for all}\ \overline{\ell}_k\right\}\ .$$

We shall write Y_g, $b\left(y,\overline{\ell}_m,g\right)$, and $b\left(\overline{\ell}_m,g\right)$ for $g\in\mathbf{g}_P$ as $Y_{\left(g_P,\overline{a}_Z\right)}$, $b\left(y,\overline{\ell}_m,\{g_P,\overline{a}_Z\}\right)$, and $b\left(\overline{\ell}_m,\{g_P,\overline{a}_Z\}\right)$ in obvious notation where $b\left(\overline{\ell}_m,g\right)$ is defined in Appendix 2. Thus $Y_{\left(g_P,\overline{a}_Z\right)}$ is the outcome Y in a hypothetical study in which the treatment $\overline{a}_Z\equiv\overline{a}_{ZK}$ is assigned non-dynamically, but $\overline{a}_P$ may be assigned dynamically. Note that the function $g_P\left(\overline{\ell}_k\right)\equiv\left\{g_P\left(t_m,\overline{\ell}_m\right);m=0,\ldots,k\right\}$ are functions taking values in the support $\overline{\mathbf{a}}_{Pk}$ of $\overline{a}_{Pk}$.

The *direct-effect g-null and g-null mean hypotheses* of no direct effect of $\overline{a}_P$ controlling for $\overline{a}_Z$ are, respectively,

$$pr\left[Y_{\left(g_{P1},\overline{a}_Z\right)}>y\right]=pr\left[Y_{\left(g_{P2},\overline{a}_Z\right)}>y\right]\qquad\text{(A3.1)}$$

and

$$E\left[Y_{\left(g_{P1},\overline{a}_Z\right)}-Y_{\left(g_{P2},\overline{a}_Z\right)}\right]=0\qquad\text{(A3.2)}$$

for all $\overline{a}_Z\equiv\overline{a}_{ZK},g_{P1},g_{P2}$. Under Assumptions (3.1), (3.3), and (3.10), (A3.1) and (A3.2) are, respectively, equivalent to the direct effect "g"-null and "g"-null mean hypotheses

$$b\left(y,\{g_{P1},\overline{a}_Z\}\right)=b\left(y,\{g_{P2},\overline{a}_Z\}\right)\qquad\text{(A3.3)}$$

and

$$b\left(\{g_{P1},\overline{a}_Z\}\right)=b\left(\{g_{P2},\overline{a}_Z\}\right)\ .\qquad\text{(A3.4)}$$

Direct Effect SNMMs:

We shall first study direct-effect SNMMs and direct-effect pseudo-SNMMs. To describe these models, set $t_k\left(\overline{\ell}_k,\overline{a}_k\right)=\gamma\left(\overline{\ell}_k,\overline{a}_k,F\right)$ with the right hand side as defined in Appendix 2. For $m=k,\ldots,1$, define $t_k\left(\overline{\ell}_{m-1},\overline{a}_k\right)=$
$E\left[t_k\left(\left\{\overline{\ell}_{m-1},L_m\right\},\overline{a}_k\right)\mid\overline{L}_{m-1}=\overline{\ell}_{m-1},\overline{A}_{m-1}=\overline{a}_{m-1}\right]$. Let $t_k\left(\overline{a}_k\right)\equiv t_k\left(\overline{\ell}_{-1},\overline{a}_k\right)\equiv E\left[t_k\left(L_0,\overline{a}_k\right)\right]$. Set $r_k\left(\overline{\ell}_m,\overline{a}_k\right)=t_k\left(\overline{\ell}_m,\overline{a}_k\right)-t_k\left(\overline{\ell}_{m-1},\overline{a}_k\right)$. Hence,

$$\gamma\left(\overline{\ell}_k,\overline{a}_k,F\right)=\sum_{m=0}^{k}r_k\left(\overline{\ell}_m,\overline{a}_k\right)+t_k\left(\overline{a}_k\right)\qquad\text{(A3.5)}$$

where, by construction,

$$E\left[r_k\left(\overline{L}_m,\overline{a}_k\right)\mid\overline{L}_{m-1},\overline{A}_{m-1}=\overline{a}_{m-1}\right]=0$$

and, by $\gamma\left(\overline{\ell}_k,\overline{a}_k,F\right)=0$ when $a_k=0$,

$$r_k\left(\overline{\ell}_m,\overline{a}_k\right)=t_k\left(\overline{a}_k\right)=0 \text{ when } a_k=0\ .$$

Now set $r_k^*\left(\overline{\ell}_m,\overline{a}_k\right)\equiv r_k\left(\overline{\ell}_m,\overline{a}_k\right)-r_k\left(\left\{\overline{\ell}_{m-1},\ell_m=0\right\},\overline{a}_k\right)$, so that

$$r_k\left(\overline{\ell}_m,\overline{a}_k\right)=r_k^*\left(\overline{\ell}_m,\overline{a}_k\right)-E\left[r_k^*\left(\left\{\overline{\ell}_{m-1},L_m\right\},\overline{a}_k\right)\mid\overline{L}_{m-1}=\overline{\ell}_{m-1},\overline{A}_{m-1}=\overline{a}_{m-1}\right] \qquad \text{(A3.5a)}$$

and

$$r_k^*\left(\overline{\ell}_m,\overline{a}_k\right)=0 \text{ if } a_k=0 \text{ or } \ell_m=0\ . \qquad \text{(A3.6)}$$

Thus

$$\gamma\left(\overline{\ell}_k,\overline{a}_k,F\right)= \qquad \text{(A3.7)}$$

$$\sum_{m=0}^{k} r_k^*\left(\overline{\ell}_m,\overline{a}_k\right)-E\left[r_k^*\left(\left\{\overline{\ell}_{m-1},L_m\right\},\overline{a}_k\right)\mid\overline{L}_{m-1}=\overline{\ell}_{m-1},\overline{A}_{m-1}=\overline{a}_{m-1}\right]+t_k\left(\overline{a}_k\right)$$

If the mean of Y is unrestricted, (A3.7) is an unrestricted parameterization of $\gamma\left(\overline{\ell}_k,\overline{a}_k,F\right)$ in the sense that given any function $t_k\left(\overline{a}_k\right)$ `that` is zero when $a_k=0$, any function $r_k^*\left(\overline{\ell}_m,\overline{a}_k\right)$ satisfying (A3.6), and any law of $\overline{X}_K=\left(\overline{L}_K,\overline{A}_K\right)$, the function $\gamma\left(\overline{\ell}_k,\overline{a}_k\right)$ defined by the R.H.S. of (A3.7) will equal $\gamma\left(\overline{\ell}_k,\overline{a}_k,F\right)$ when the law of $\left(Y,\overline{X}_K\right)$ is generated as described following Eq. (A2.5); furthermore, all possible distributions of $\left(Y,\overline{X}_K\right)$ can be constructed in this manner.

An alternative interesting unrestricted parameterization is as follows. Let $b\left(\overline{\ell}_m,\overline{a}\right)$ be shorthand for $b\left(\overline{\ell}_m,g=\left(\overline{a}\right)\right)$ as defined in Appendix 2. Now define $v\left(\overline{\ell}_m,\overline{a}\right)=b\left(\overline{\ell}_m,\overline{a}\right)-b\left(\overline{\ell}_{m-1},\overline{a}\right)$ and $v^*\left(\overline{\ell}_m,\overline{a}\right)=v\left(\overline{\ell}_m,\overline{a}\right)-v\left(\left\{\overline{\ell}_{m-1},\ell_m=0\right\},\overline{a}\right)$ so that

$$v^*\left(\overline{\ell}_{m-1},\ell_m=0,\overline{a}\right)=0 \qquad \text{(A3.8)}$$

Then, we have

$$b\left(\overline{\ell}_K,\overline{a}\right)= \qquad \text{(A3.9)}$$

$$\sum_{m=0}^{K} v^*\left(\overline{\ell}_m,\overline{a}\right)-E\left[v^*\left(\left\{\overline{\ell}_{m-1},L_m\right\},\overline{a}\right)\mid\overline{L}_{m-1}=\overline{\ell}_{m-1},\overline{A}_{m-1}=\overline{a}_{m-1}\right]+b\left(\bar{a}\right)$$

(A3.9) is an unrestricted parameterization in the sense that given any law for $\left(\overline{L}_K,\overline{A}_K\right)$, any random variable σ satisfying (A2.5), any function $b\left(\overline{a}\right)$, and any function $v^*\left(\overline{\ell}_m,\overline{a}\right)$ satisfying (A3.8), if we set $Y=b\left(\overline{L}_K,\overline{A}_K\right)+\sigma$ with $b\left(\overline{L}_K,\overline{A}_K\right)$ computed by the RHS of (A3.9), then the joint distribution of $\left(Y,\overline{L}_K,\overline{A}_K\right)$ will be consistent with our choices of $v^*\left(\overline{\ell}_m,\overline{a}\right)$ and $b\left(\overline{a}\right)$. Further, all possible distributions of $\left(Y,\overline{L}_K,\overline{A}_K\right)$ can be constructed in this manner.

The relationship between the two parameterizations is given by

$$t_k\left(\overline{a}_k\right)=b\left(\left(\overline{a}_k,0\right)\right)-b\left(\left(\overline{a}_{k-1},0\right)\right) \ \textit{and}\ r_k^*\left(\overline{\ell}_m,\overline{a}_k\right)=v^*\left(\overline{\ell}_m,\left(\overline{a}_k,0\right)\right)-v^*\left(\overline{\ell}_m,\left(\overline{a}_{k-1},0\right)\right)\ .$$

If Y is a positive random variable, an analogous multiplicative version of both parameterizations can be obtained.

Our main theorem is the following.

Theorem A3.1: The following are equivalent: *(i)* The direct effect "g"-null mean hypothesis (A3.4) holds;
(ii) $t_k\left(\overline{\ell}_m, \overline{a}_k\right)$ does not depend on $(a_{Pm}, \ldots, a_{Pk})$;
(iii) $t_k\left(\overline{a}_k\right) = t_k\left(\overline{a}_{Zk}\right)$ does not depend on $\overline{a}_{Pk}$ and $r_k^*\left(\overline{\ell}_m, \overline{a}_k\right) =$
$r_k^*\left(\overline{\ell}_m, \overline{a}_{m-1}, a_{Zm}, \ldots, a_{Zk}\right)$ does not depend on $(a_{Pm}, \ldots, a_{Pk})$;
(iv) $v^*\left(\overline{\ell}_m, \overline{a}\right)$ does not depend on $\left(a_{Pm}, \ldots, a_{PK}\right)$ and $b\left(\overline{a}\right)$ does not depend on $\left(\overline{a}_{PK}\right)$;
(v) $b\left(\overline{\ell}_m, \overline{a}\right)$ does not depend on $a_{Pm}, \ldots, a_{PK}$.

The proof, which we do not include, follows easily by examining the limiting version of the Monte Carlo procedure of Sec. 8.3. The following is an easy corollary of Theorem (A3.1) and Theorem (3.2).

Corollary: If (3.1), (3.3), and (3.10) hold, then the direct effect g-null mean hypothesis (A3.2) holds if and only if for $0 \leq m \leq k, 0 \leq k \leq K$

$$E\left[Y_{g=\left(\overline{a}_k,0\right)} \mid \overline{L}_m, \overline{A}_{m-1} = \overline{a}_{m-1}\right] \text{ \textit{does not depend on} } \left(a_{Pm}, \ldots, a_{Pk}\right).$$

We could create models based on either of the unrestricted parameterizations (A3.7) or (A3.9). We choose here (A3.7) because of the connection with SNMMs. We say the data follow a direct effect pseudo-SNMM for the direct effect of treatment $\overline{a}_P$ controlling for treatment $\overline{a}_Z$ if $t_k\left(\overline{a}_k\right) = t_k\left(\overline{a}_k, \psi_{Pt0}, \psi_{Zt0}\right)$ and $r_k^*\left(\overline{\ell}_m, \overline{a}_k\right) = r_k^*\left(\overline{\ell}_m, \overline{a}_k, \psi_{Pr0}, \psi_{Zr0}\right)$ where *(i)* $t_k\left(\overline{a}_k, \psi_{Pt}, \psi_{Zt}\right)$ and $r_k^*\left(\overline{\ell}_m, \overline{a}_k, \psi_{Pr}, \psi_{Zr}\right)$ are known functions that take the value zero if $a_k = 0$ or $\psi = 0$ with $\psi \equiv (\psi_P, \psi_Z)$, $\psi_P = (\psi_{Pt}, \psi_{Pr})$, $\psi_Z = (\psi_{Zt}, \psi_{Zr})$, *(ii)* $r_k^*\left(\overline{\ell}_m, \overline{a}_k, \psi_{Pr}, \psi_{Zr}\right) = 0$ if $\ell_m = 0$, *(iii)* $t_k\left(\overline{a}_k, \psi_{Pt} = 0, \psi_{Zt}\right)$ depends on $\overline{a}_k$ only through $\overline{a}_{Zk}$, *(iv)* and $r_k^*\left(\overline{\ell}_m, \overline{a}_k, \psi_{Pr} = 0, \psi_{Zr}\right)$ depends on $\overline{a}_k$ only through $\overline{a}_{m-1}, a_{Zm}, \ldots, a_{Zk}$.

Remark: There are no restrictions on the functions $r_k^*\left(\overline{\ell}_m, \overline{a}_k, \psi_r\right)$ and $t_k\left(\overline{a}_k, \psi_t\right)$ except for (i)-(iv) above. Here $\psi_r \equiv (\psi_{Pr}, \psi_{Zr})$.

Example: With ℓ_m univariate,

$$r_k^*\left(\overline{\ell}_m, \overline{a}_k, \psi_r\right) = \psi_{Pr} a_{Pk} \ell_m \left[\sum_{j=0}^{k-1} \left(a_{Pj} + a_{Zj}\right)\right] + \psi_{Zr} a_{Zk} \ell_m \left[\sum_{j=0}^{m-1} a_{Pj} + \sum_{j=0}^{k-1} a_{Zj}\right]$$

$$t\left(\overline{a}_k, \psi_t\right) = \psi_{Pt1} a_{Pk} \left(\sum_{m=0}^{k-1} a_{Pm} + a_{Zm}\right) + \psi_{Pt2} a_{Zk} \left(\sum_{m=0}^{k-1} a_{Pm}\right) + \psi_{Zt} a_{Zk} \left(\sum_{m=0}^{k-1} a_{Zm}\right).$$

It follows from the "g"-null mean theorem of Appendix 2 and Theorem (A3.1) that *(i)* $\psi_0 = 0$ if and only if the "g"-null mean hypothesis that $b\left(g_1\right) = b\left(g_2\right)$ for all g_1, g_2 holds, and *(ii)* $\psi_{P0} = 0$ if and only if the direct effects "g"-null mean hypothesis (A3.4) holds.

If (3.1), (3.3), and (3.10) hold so that $\gamma\left(\overline{\ell}_k, \overline{a}_k, F\right) = \gamma^\dagger\left(\overline{\ell}_k, \overline{a}_k\right)$, we also call our direct-effect pseudo-SNMM a direct-effect SNMM. In this case, *(i)* $\psi_0 = 0$ if and only if the g-null mean hypothesis (5.1) holds and *(ii)* $\psi_{P0} = 0$ if and only if the direct effect g-null mean hypothesis (A3.2) holds.

To estimate the parameter ψ_0 of a direct-effect pseudo-SNMM by "generalized" g-estimation, we proceed as follows. Write $\psi_r = (\psi_{Pr}, \psi_{Zr})$ and $\psi_t = (\psi_{Pt}, \psi_{Zt})$. First we estimate $f\left[\ell_m \mid \overline{\ell}_{m-1}, \overline{a}_{m-1}\right]$ by specifying a parametric model $f\left[\ell_m \mid \overline{\ell}_{m-1}, \overline{a}_{m-1}; \omega_3\right]$ depending on an unknown parameter ω_3 and find $\widehat{\omega}_3$ that maximizes $\prod_{i=1}^{n} f\left[\ell_{mi} \mid \overline{\ell}_{(m-1)\,i}, \overline{a}_{(m-1)\,i}; \omega_3\right]$. We then construct $r_k\left(\overline{\ell}_m, \overline{a}_k; \psi_r, \widehat{\omega}_3\right)$ by evaluating Eq. (A3.5a) at $\widehat{\omega}_3$ and $r_k^*\left(\overline{\ell}_m, \overline{a}_k; \psi_r\right)$. Next we use $r_k\left(\overline{\ell}_m, \overline{a}_k; \psi_r, \widehat{\omega}_3\right)$ and $t_k\left(\overline{a}_k; \psi_t\right)$ to obtain $\gamma\left(\overline{\ell}_k, \overline{a}_k; \psi, \widehat{\omega}_3\right)$ by Eq. (A3.5). We next compute $H\left(\psi, \widehat{\omega}_3\right) = Y - \sum_{m=0}^{K} \gamma\left(\overline{L}_m, \overline{A}_m; \psi, \widehat{\omega}_3\right)$. Finally we estimate the unknown parameter ψ by g-estimation as in Sec. 8.3 with $H\left(\psi, \widehat{\omega}_3\right)$ replacing $H\left(\psi\right)$.

To estimate ψ by maximum likelihood, we would proceed as in Appendix 2 in the paragraph following Eq. (A2.9), except that, in computing the residual $\sigma\left(\psi,\omega\right)$, *(i)* $H\left(\psi\right)$ is replaced by $H\left(\psi,\omega_3\right)$ and *(ii)* all η's are replaced by ω's in (A2.7).

Direct-Effect SNDMs:

We now describe direct-effect SNDMs and pseudo-SNDMs. We first provide the notation necessary to describe our models. Write $b\left(y,\overline{\ell}_k,\overline{a}\right)$ and $b\left(y,\overline{a}\right)$ as shorthand for $b\left(y,\overline{\ell}_k,g=\left(\overline{a}\right)\right)$ and $b\left(y,g=\left(\overline{a}\right)\right)$. Note $b\left(y,\overline{\ell}_{m-1},\overline{a}\right)=$
$E\left[b\left(y,\left(\overline{\ell}_{m-1},L_m\right),\overline{a}\right)\mid\overline{L}_{m-1}=\overline{\ell}_{m-1},\overline{A}_{m-1}=\overline{a}_{m-1}\right]$. Under (3.1), (3.3), and the sequential randomization assumption (3.10), Theorem (3.2) implies
$b\left(y,\overline{\ell}_m,\left(\overline{a}_k,0\right)\right)=pr\left[Y_{g=\left(\overline{a}_k,0\right)}>y\mid\overline{L}_m=\overline{\ell}_m,\overline{A}_{m-1}=\overline{a}_{m-1}\right]$. Now define $t_k\left(y,\overline{\ell}_m,\overline{a}_k\right)$ for $m=-1,0,\ldots,k$ by

$$b\left[y,\overline{\ell}_m,\left(\overline{a}_k,0\right)\right]=b\left[t_k\left(y,\overline{\ell}_m,\overline{a}_k\right),\overline{\ell}_m,\left(\overline{a}_{k-1},0\right)\right]$$

so, in particular, $t_k\left(y,\overline{\ell}_k,\overline{a}_k\right)=\gamma\left(y,\overline{\ell}_k,\overline{a}_k,F\right)$. Hence, under (3.1), (3.3), and (3.10), $t_k\left(y,\overline{\ell}_m,\overline{a}_k\right)$ satisfies

$$\begin{gathered}pr\left[Y_{g=\left(\overline{a}_k,0\right)}>y\mid\overline{L}_m=\overline{\ell}_m,\overline{A}_{m-1}=\overline{a}_{m-1}\right]=\\ pr\left[Y_{g=\left(\overline{a}_{k-1},0\right)}>t_k\left(y,\overline{\ell}_m,\overline{a}_k\right)\mid\overline{L}_m=\overline{\ell}_m,\overline{A}_{m-1}=\overline{a}_{m-1}\right].\end{gathered}$$

In particular, $t_k\left(y,\overline{\ell}_k,\overline{a}_k\right)=\gamma^{\dagger}\left(y,\overline{\ell}_k,\overline{a}_k\right)$. Our main results are the following theorem and corollary.

Theorem (A3.2): The following are equivalent: *(i)* The direct effect "g"-null hypothesis (A3.3) holds; *(ii)* for $0\leq m\leq k$, $0\leq k\leq K$,

$$t_k\left(y,\overline{\ell}_m,\overline{a}_k\right)\ \textit{does not depend on}\ \left(a_{Pm},\ldots,a_{Pk}\right)\ ; \tag{A3.10}$$

$$\textit{(iii)}\ b\left(y,\overline{\ell}_k,\overline{a}\right)\ \textit{does not depend on}\ \left(a_{Pk},\ldots,a_{PK}\right),0\leq k\leq K \tag{A3.11}$$

Corollary (A3.1): If (3.1), (3.3), and (3.10) hold, the following are equivalent: the direct-effect g-null hypothesis (A3.1) holds; *(ii)* (A3.10) is true; *(iii)* (A3.11) is true.

The proof of Theorem A3.2 which we do not give follows easily from the limiting form of the Monte Carlo algorithm of Sec. 7.5. Now define $v\left(y,\overline{\ell}_m,\overline{a}\right)$ by

$$b\left(y,\overline{\ell}_m,\overline{a}\right)=b\left\{v\left(y,\overline{\ell}_m,\overline{a}\right),\overline{\ell}_{m-1},\overline{a}\right\}\ . \tag{A3.12}$$

Now let

$$v^*\left(y,\overline{\ell}_m,\overline{a}\right)\equiv v\left(y,\overline{\ell}_m,\overline{a}\right)-v\left(y,\left\{\overline{\ell}_{m-1},\ell_m=0\right\},\overline{a}\right)\ , \tag{A3.12a}$$

so

$$v^*\left(y,\overline{\ell}_m,\overline{a}\right)=0\ \textit{if}\ \ell_m=0\ . \tag{A3.13}$$

We have from these definitions that

$$\begin{gathered}E[b\{\left[v^*\left(y,\overline{L}_m,\overline{a}\right)+v\left(y,\left\{\overline{\ell}_{m-1},\ell_m=0\right\},\overline{a}\right)\right],\overline{\ell}_{m-1},\overline{a}\}\mid\\ \overline{L}_{m-1}=\overline{\ell}_{m-1},\overline{A}_{m-1}=\overline{a}_{m-1}]=b\left(y,\overline{\ell}_{m-1},\overline{a}\right)\ .\end{gathered} \tag{A3.14}$$

Now $b(y,\overline{a})$ and $v^*\left(y,\overline{\ell}_m,\overline{a}\right)$ constitute an unrestricted parameterization in the sense that any survivor function $b(y,\overline{a})$, function $v^*\left(y,\overline{\ell}_m,\overline{a}\right)$ satisfying (A3.13), and law for $\left(\overline{L}_K,\overline{A}_K\right)$ determine a unique law F for $\left(Y,\overline{L}_K,\overline{A}_K\right)$ by $pr\left[Y>y \mid \overline{L}_K,\overline{A}\right]=b\left(y,\overline{L}_K,\overline{A}\right)$ with $\overline{A}\equiv\overline{A}_K$ and $b\left(y,\overline{\ell}_K,\overline{a}\right)$ determined recursively from $b(y,\overline{a})\equiv b\left(y,\overline{\ell}_{-1},\overline{a}\right)$, $v^*\left(y,\overline{\ell}_m,\overline{a}\right)$, (A3.12) and (A3.14). Specifically, given $b\left(y,\overline{\ell}_{m-1},\overline{a}\right)$ and $v^*\left(y,\overline{\ell}_m,\overline{a}\right)$, we solve (A3.14) for $v\left(y,\left\{\overline{\ell}_{m-1},\ell_m=0\right\},\overline{a}\right)$, and then compute $v\left(y,\overline{\ell}_m,\overline{a}\right)$ by (A3.12a). Finally this allows us to compute $b\left(y,\overline{\ell}_m,\overline{a}\right)$ from (A3.12). Furthermore, any law F for $\left(Y,\overline{L}_K,\overline{A}_k\right)$ can be obtained by this method.

Further, the "g"-null hypothesis (5.2) holds if and only if both $b(y,\overline{a})$ does not depend on $\overline{a}$ and $v^*\left(y,\overline{\ell}_m,\overline{a}\right)$ does not depend on $a_m,\ldots,a_K$. Further, the direct-effect "g"-null hypothesis holds if and only if *(i)* $b(y,\overline{a})$ does not depend on $\overline{a}_{PK}$ and *(ii)* $v^*\left(y,\overline{\ell}_m,\overline{a}\right)$ does not depend on $a_{Pm},\ldots,a_{PK}$. Thus, likelihood-based inference can be carried out by specifying parametric models for $b(y,\overline{a})$ and $v^*\left(y,\overline{\ell}_m,\overline{a}\right)$ (satisfying (A3.13)) with the parameters ψ_P and ψ_Z such that $\psi_P=0$ implies *(i)* and *(ii)* just above. This parameterization is not closely related to SNDM models. We now describe another unrestricted parameterization that allows a connection to SNDM models.

Now for $m=0,\ldots,k$, define $r_k\left(y,\overline{\ell}_m,\overline{a}_k\right)$ by $r_k\left[t_k\left(y,\overline{\ell}_{m-1},\overline{a}_k\right),\overline{\ell}_m,\overline{a}_k\right]=t_k\left(y,\overline{\ell}_m,\overline{a}_k\right)$. Then the following lemma is easy to prove.

Lemma A3.1: $r_k\left(y,\overline{\ell}_m,\overline{a}_k\right)$ satisfies

$$E[b\left\{r_k\left(y,\overline{L}_m,\overline{a}_k\right),\overline{L}_m,(\overline{a}_{k-1},0)\right\} \mid \overline{L}_{m-1}=\overline{\ell}_{m-1},\overline{A}_{m-1}=\overline{a}_{m-1}]=b\left(y,\overline{\ell}_{m-1},(\overline{a}_{k-1},0)\right).$$

Lemma (A3.1) has the following Corollary.

Corollary (A3.2): If (3.1), (3.3), and (3.10) hold, $r_k\left(y,\overline{\ell}_m,\overline{a}_k\right)$ satisfies

$$E\{pr\left[Y_{g=(\overline{a}_{k-1},0)}>r_k\left(y,\overline{L}_m,\overline{a}_k\right) \mid \overline{L}_m,\overline{A}_{m-1}=\overline{a}_{m-1}\right] \mid$$
$$\overline{L}_{m-1}=\overline{\ell}_{m-1},\overline{A}_{m-1}=\overline{a}_{m-1}\}=pr\left[Y_{g=(\overline{a}_{k-1},0)}>y \mid \overline{L}_{m-1}=\overline{\ell}_{m-1},\overline{A}_{m-1}=\overline{a}_{m-1}\right].$$

Now define $r_k^*\left(y,\overline{\ell}_m,\overline{a}_k\right)\equiv r_k\left(y,\overline{\ell}_m,\overline{a}_k\right)-r_k\left(y,\left\{\overline{\ell}_{m-1},\ell_m=0\right\},\overline{a}_k\right)$ so that $r_k^*\left(y,\overline{\ell}_m,\overline{a}_k\right)=0$ if $\ell_m=0$ or $a_k=0$. We now have the following easy corollary to Theorem (A3.2). Set $t_k\left(y,\overline{a}_k\right)\equiv t_k\left(y,\overline{\ell}_{-1},\overline{a}_k\right)$.

Corollary (A3.3): The direct-effect "g"-null hypothesis (A3.3) holds if and only if for all k and $m\leq k$, $r_k^*\left(y,\overline{\ell}_m,\overline{a}_k\right)=r_k^*\left(y,\overline{\ell}_m,\overline{a}_{m-1},a_{Zm},\ldots,a_{Zk}\right)$ does not depend on $\left(a_{Pm},\ldots,a_{Pk}\right)$ and $t_k\left(y,\overline{a}_k\right)=t_k\left(y,\overline{a}_{Zk}\right)$ does not depend on $\overline{a}_{Pk}$.

We say the data follow a direct effect pseudo-SNDM for the direct effect of treatment $\overline{a}_P$ controlling for treatment $\overline{a}_Z$ if $t_k\left(y,\overline{a}_k\right)=t_k\left(y,\overline{a}_k,\psi_{Pt0},\psi_{Zt0}\right)$ and $r_k^*\left(y,\overline{\ell}_m,\overline{a}_k\right)=r_k^*\left(y,\overline{\ell}_m,\overline{a}_k,\psi_{Pr0},\psi_{Zr0}\right)$ where *(i)* $t_k\left(y,\overline{a}_k,\psi_{Pt},\psi_{Zt}\right)$ and $r_k^*\left(y,\overline{\ell}_m,\overline{a}_k,\psi_{Pr},\psi_{Zr}\right)$ are known functions that take the values y or zero respectively if $a_k=0$ or $\psi=0$ with $\psi\equiv\left(\psi_P,\psi_Z\right)$, $\psi_P=\left(\psi_{Pt},\psi_{Pr}\right)$, $\psi_Z=\left(\psi_{Zt},\psi_{Zr}\right)$, *(ii)* $r_k^*\left(y,\overline{\ell}_m,\overline{a}_k,\psi_{Pr},\psi_{Zr}\right)=0$ if $\ell_m=0$, *(iii)* $t_k\left(y,\overline{a}_k,\psi_{Pt}=0,\psi_{Zt}\right)$ depends on $\overline{a}_k$ only through $\overline{a}_{Zk}$, *(iv)* and $r_k^*\left(\gamma,\overline{\ell}_m,\overline{a}_k,\psi_{Pr}=0,\psi_{Zr}\right)$ depends on $\overline{a}_k$ only through $\overline{a}_{m-1},a_{Zm},\ldots,a_{Zk}$.

It follows from the "g"-null theorem and Corollary (A3.3) that *(i)* $\psi_0=0$ if and only if the "g"-null hypothesis (5.2) holds, and *(ii)* $\psi_{P0}=0$ if and only if the direct effect "g"-null hypothesis (A3.3) holds.

If (3.1), (3.3), and (3.10) hold so that $\gamma\left(y,\overline{\ell}_k,\overline{a}_k,F\right)=\gamma^\dagger\left(y,\overline{\ell}_k,\overline{a}_k\right)$, we also call our direct-effect pseudo-SNDM a direct-effect SNDM. In this case, *(i)* $\psi_0=0$ if and only if the g-null hypothesis (5.1) holds and *(ii)* $\psi_{P0}=0$ if and only if the direct effect g-null hypothesis (A3.1) holds.

We now consider estimation of the parameter ψ_0 for a direct-effect pseudo-SNDM with maximum likelihood. We do not have a simple generalization of g-estimation. As in Sec. 7, let $f\left(\ell_m \mid \overline{\ell}_{m-1},\overline{a}_{m-1},h;\phi\right)$

be a model indexed by ϕ for $L_m \mid \overline{L}_{m-1}, \overline{A}_{m-1}, H(\gamma)$ and $f(h;\eta)$ be a model indexed by η for the marginal law of $H(\gamma)$. Now let $\theta = (\psi, \phi, \eta)$ where ψ is the parameter vector for the direct-effect pseudo-SNMM and let F_θ be a law of the observed data generated under parameter θ. Since our parameterization can be shown to be unrestricted, the above models need not be restricted. In addition, our pseudo-SNDM model is unrestricted except as in *(i)-(iv)* of its definition.

Now using the fact that $H(\gamma) \coprod A_m \mid \overline{L}_m, \overline{A}_{m-1}$, Lemma A3.1 can be rewritten as follows after some algebraic manipulation.

$$\begin{gathered}\Gamma_k\left(\overline{\ell}_{m-1}, \overline{a}_{k-1}, y; F_\theta\right) = \\ \Gamma_k\left[\overline{\ell}_{m-1}, \overline{a}_{k-1}, r_k^*\left(y, \overline{\ell}_m, \overline{a}_k; \psi\right) + r_k\left(y, \left\{\overline{\ell}_{m-1}, \ell_m = 0\right\}, \overline{a}_k; F_\theta\right)\right]\end{gathered} \tag{A3.15}$$

where, for any function $c\left(y, \overline{\ell}_m, \overline{a}_k\right)$, $\Gamma_k\left(\overline{\ell}_{m-1}, \overline{a}_{k-1}, c\left(y, \overline{\ell}_m, \overline{a}_k\right); F_\theta\right) \equiv$
$\int \cdots \int I\left\{q\left[h, \overline{\ell}_{k-1}, \overline{a}_{k-1}; F_\theta\right] > c\left(y, \overline{\ell}_m, \overline{a}_k\right)\right\} \prod_{j=0}^{k-1} f\left[\ell_j \mid \overline{\ell}_{j-1}, \overline{a}_{j-1}, h; \phi\right]$
$\prod_{j=m}^{k-1} d\mu\left(\ell_j\right) dF(h;\eta) / \int \prod_{j=0}^{m-1} f\left[\ell_j \mid \overline{\ell}_{j-1}, \overline{a}_{j-1}, h; \phi\right] dF(h;\eta)$. The key identity in deriving (A3.15) is the fact that $H(\gamma) \coprod A_m \mid \overline{L}_m, \overline{A}_{m-1}$ implies $f\left(h \mid \overline{\ell}_k, \overline{a}_{k-1}\right) = \prod_{j=0}^{k} f\left(\ell_j \mid \overline{\ell}_{j-1}, \overline{a}_{j-1}, h\right) f(h) /$
$\int \prod_{j=0}^{k} f\left(\ell_j \mid \overline{\ell}_{j-1}, \overline{a}_{j-1}, h\right) f(h)\, dh$ where h represents realizations of $H(\gamma)$.

Now, as in Sec. 7.5, we note $q\left(h, \overline{\ell}_k, \overline{a}_k; F_\theta\right)$ depends on F_θ through $\gamma^{-1}\left(y, \overline{\ell}_m, \overline{a}_m, F_\theta\right)$ for $m \leq k$, and, in our model, $\gamma\left(y, \overline{\ell}_k, \overline{a}_k; F_\theta\right) \equiv t_k\left(y, \overline{\ell}_k, \overline{a}_k; F_\theta\right)$ is obtained recursively by $t_k\left(y, \overline{\ell}_m, \overline{a}_k; F_\theta\right) =$
$r_k\left[t_k\left(y, \overline{\ell}_{m-1}, \overline{a}_k; F_\theta\right); \overline{\ell}_m, \overline{a}_k; F_\theta\right]$ with $t_k\left(y, \overline{\ell}_{-1}, \overline{a}_k; F_\theta\right) \equiv t_k\left(y, \overline{a}_k; \psi\right)$ and $r_k\left(y, \overline{\ell}_m, \overline{a}_k; F_\theta\right) =$
$r_k^*\left(y, \overline{\ell}_m, \overline{a}_k; \psi\right) + r_k\left(y, \left\{\overline{\ell}_{m-1}, \ell_m = 0\right\}, \overline{a}_k; F_\theta\right)$. Thus, given $\theta = (\psi, \phi, \eta)$ and
$r_j\left[y, \left\{\overline{\ell}_{m-1}, \ell_m = 0\right\}, \overline{a}_j; F_\theta\right], 0 \leq j \leq k-1$ [and thus $q\left(h, \overline{\ell}_{k-1}, \overline{a}_{k-1}; F_\theta\right)$] are known, it follows that all terms in (A3.15) [including $q\left(h, \overline{\ell}_{k-1}, \overline{a}_{k-1}; F_\theta\right)$] are known except for $r_k\left[y, \left\{\overline{\ell}_{m-1}, \ell_m = 0\right\}, \overline{a}_k; F_\theta\right]$
which can therefore be solved for recursively. In general, $r_k\left[y, \left\{\overline{\ell}_{m-1}, \ell_m = 0\right\}, \overline{a}_k; F_\theta\right]$ and thus
$\gamma\left(y, \overline{\ell}_k, \overline{a}_k; F_\theta\right)$ and $H(\gamma)$ will be functions of all the parameters ψ, ϕ, and η of θ. Thus write
$r_k\left(y, \left\{\overline{\ell}_{m-1}, \ell_m = 0\right\}, \overline{a}_k; \theta\right)$ and $H(\theta) = H(\psi, \phi, \eta)$ to emphasize the functional dependence on all of θ. The maximum likelihood estimator of $\theta = (\psi, \phi, \eta)$ is obtained by maximizing Eq. (7.8) with $h_i(\psi)$ replaced by the realization $h_i(\theta)$ of $H(\theta)$. Additional study of the computation problems involved will be useful.

Finally, Robins and Wasserman (1996) provide another unrestricted parameterization for the law of $\left(Y, \overline{L}_K, \overline{A}_K\right)$ based on multiplicative models for the hazard functions of the survivor functions $b(y, \overline{a})$ and $b\left(y, \overline{\ell}_k, \overline{a}\right)$. The models contain a parameter ψ_P that is zero under the direct-effect "g"-null hypothesis.

REFERENCES

Arjas, E. (1989). Survival models and martingale dynamics (with discussion). *Scandinavian Journal of Statistics*, 15, 177-225.

Gill, R. and Robins, J. (1996). Some measure theoretic aspects of causal models. (In preparation.)

Heckerman, D. and Shachter, R. (1995). Decision-theoretic foundations for causal inference. *Journal of Artificial Intelligence Research*, 3, 405-430.

Holland, P. (1989), Reader reaction: Confounding in epidemiologic studies, *Biometrics*, 45, 1310-1316.

Lewis, D.K. (1973), **Counterfactuals**. Cambridge: Harvard University Press.

Loomis, B. and Sternberg, S. (1968). **Advanced Calculus**. Addison Wesley.

Pearl, J. (1995), Causal diagrams for empirical research, *Biometrika*, 82, 669-690.

Pearl, J. and Robins, J.M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables, From: **Uncertainty in Artificial Intelligence:** Proceedings of the Eleventh Conference on Artificial Intelligence, August 18-20, 1995, McGill University, Montreal, Quebec, Canada. San Francisco, CA: Morgan Kaufmann. pp. 444-453.

Pearl, J. and Verma, T. (1991). A Theory of Inferred Causation. In: **Principles of Knowledge, Representation and Reasoning: Proceedings of the Second International Conference.** (Eds. J.A. Allen, R. Fikes, and E. Sandewall). 441-452.

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. (1987), Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect", *Computers and Mathematics with Applications*, 14, 923-945.

Robins, J.M. (1989), The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies, In: **Health Service Research Methodology: A Focus on AIDS**, eds. Sechrest, L., Freeman, H., Mulley, A., NCHSR, U.S. Public Health Service, 113-159.

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. (1993), Analytic methods for estimating HIV-treatment and cofactor effects,In: **Methodological Issues in AIDS Mental Health Research**, eds. Ostrow, D.G., and Kessler, R.C., NY: Plenum Press, 213-290.

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. (1995a). Estimating the Causal Effect of a Time-varying Treatment on Survival using Structural Nested Failure Time Models,(To appear, *Statistica Neederlandica*).

Robins, J.M. (1995b). Discussion of "Causal Diagrams for empirical research" by J. Pearl,*Biometrika*, 82, 695-698.

Robins, J.M. (1998). Correcting for non-compliance in bioequivalence trials,*Statistics in Medicine* (To appear).

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

Robins, J.M. and Pearl, J. (1996). Causal effects of dynamic policies.In preparation.

Robins, J.M. and Wasserman, L. (1996). Parameterizations of directed acyclic graphs for the estimation of overall and direct effects of multiple treatments.Technical Report, Department of Epidemiology, Harvard School of Public Health.

Rosenbaum, P.R. (1984), Conditional permutation tests and the propensity score in observational studies,*Journal of the American Statistical Association*, 79, 565-574.

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

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

Rubin, D.B. (1978), Bayesian inference for causal effects: The role of randomization,*The Annals of Statistics*, 6, 34-58.

Spirtes, P., Glymour, C., and Scheines, R. (1993). **Causation, Prediction, and Search**. New York: Springer Verlag.

Acknowledgements: Financial support for this research was provided by NIH Grant AI32475.
Richard Gill, Larry Wasserman, and Sander Greenland provided valuable comments.