To appear, Statistical Models in Epidemiology: The Environment and Clinical Trials.
M.E. Halloran and D. Berry, Editors, NY: Springer-Verlag, 1999.

Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference

James M. Robins
Professor of Epidemiology and Biostatistics
Harvard School of Public Health
677 Huntington Avenue
Boston, MA 02115
robins@hsph.harvard.edu

Abstract: Robins (1993, 1994, 1997, 1998ab) has developed a set of causal or counterfactual models, the structural nested models (SNMs). This paper describes an alternative new class of causal models- the (non-nested) marginal structural models (MSMs). We will then describe a class of semiparametric estimators for the parameters of these new models under a sequential randomization (i.e., ignorability) assumption. We then compare the strengths and weaknesses of MSMs versus SNMs for causal inference from complex longitudinal data with time-dependent treatments and confounders. Our results provide an extension to continuous treatments of propensity score estimators of an average treatment effect.

1 Introduction

Robins (1993, 1994, 1997, 1998ab) has developed a set of causal or counterfactual models, the structural nested models (SNMs). Robins (1998abcd) has recently described an alternative new class of causal models – the (non-nested) marginal structural models (MSMs). We describe a class of semiparametric estimators for the parameters of these new models under a sequential randomization (i.e., ignorability) assumption. We then compare the strengths and weaknesses of MSMs versus SNMs for causal inference from complex longitudinal data with time-dependent treatments and confounders. Two major strengths of MSMs compared to SNMs are as follows.

- MSMs can be used to provide semiparametric estimates of the causal effect of a time-dependent treatment on a binary outcome using models (e.g. logistic models) which naturally respect the fact that probabilities lie in the interval $[0, 1]$.

- MSMs cohere much more closely than do SNMs with models for the analysis of time-dependent treatments that are standardly used in the absence of time-dependent confounders. For example, in the absence of time-dependent confounders, a time-dependent Cox proportional hazards model for the effect of time-dependent treatment on a time-to-event (survival time) outcome is commonly employed. The MSMs provide a natural extension of the time-dependent proportional hazards model. Unlike the usual time-dependent Cox model, the marginal structural time-dependent Cox model can be used to obtain valid causal inferences for the effect of a time-varying treatment in the presence of time-varying confounding factors. [We remind the reader that, as discussed in Robins (1986), one cannot estimate the effect of a time-dependent treatment on survival in the presence of time-dependent confounding factors by using an ordinary time-dependent Cox model that adjusts for the time-dependent confounding factors since, in general, these time-dependent confounding factors will be both determinants of later treatment and affected by earlier treatment. Marginal structural Cox models overcome this deficiency.] Disadvantages of MSMs are discussed later. The relationship of our approach to the propensity score approach of Rosenbaum and Rubin (1983) is considered in Section 4.1.

We now give a somewhat informal introduction to marginal structural models, and we report the results of a preliminary data analysis of AIDS Clinical Trial Group (ACTG) Trial 002 using MSMs. We begin with the following simple setting. Consider a study of AIDS patients. Let $A(t)$ be the dose of a treatment of interest, say AZT, at time t with time measured as days since start of follow-up. Let Y be an outcome of interest measured at end-of-follow-up at time $K+1$. Our goal is to estimate the causal effect of the time-dependent treatment $A(t)$ on the mean of Y. Let $\overline{A}(t) = \{A(u); 0 \leq u \leq t\}$ be treatment history through t and let $\overline{L}(t) = \{L(u); 0 \leq u \leq t\}$ be the history through t of a vector of relevant prognostic factors $L(u)$ for (i.e., predictors of) Y, such as CD4 lymphocyte count, white blood count (WBC), hematocrit, age, gender, etc. Suppose Y is a dichotomous outcome (e.g., $Y = 1$ if HIV RNA is detectable in the blood and zero otherwise), and we entertain a model that says the mean of Y given AZT history, $\overline{A} \equiv \overline{A}(K+1)$, is a linear logistic function of a subject's cumulative AZT dose. We write the model

$$E\left[Y \mid \overline{A}\right] = g\left(\overline{A}; \gamma\right)$$

where

$$g\left(\overline{A}; \gamma\right) = \left[1 + \exp\left\{-\gamma_1 - \gamma_2 cum\left(\overline{A}\right)\right\}\right]^{-1}$$

and $cum\left(\overline{A}\right) = \int_0^{K+1} A(t)\, dt$ is the subject's cumulative treatment. The maximum likelihood estimator (MLE) of γ can then be computed from the observed data $O_i = \left(\overline{L}_i, \overline{A}_i, Y_i\right), i = 1, \ldots, n$, on the n study subjects using standard logistic regression software with Y as the Bernoulli outcome variable and $cum\left(\overline{A}\right)$ as the regressor. That is, the MLE of $\gamma = (\gamma_1, \gamma_2)'$ maximizes ${}_{i=1}^{n} Lik_i(\gamma)$ with $Lik_i(\gamma) = g\left(\overline{A}_i; \gamma\right)^{Y_i}\left[1 - g\left(\overline{A}_i; \gamma\right)\right]^{1-Y_i}$ being the likelihood contribution for a single subject. [Note $Lik_i(\gamma)$ does not depend on the patient's prognostic factor history $\overline{L}_i \equiv \overline{L}_i(K+1)$.] Alternatively, we could have used Bayesian methods by specifying a prior distribution for γ and then estimating γ by its posterior mean given the data. In reasonable large samples, the MLE and Bayes estimate will closely approximate one another.

Causal Interpretation of Regression Parameters: The question then is when does γ_2 have an interpretation as the causal effect of treatment history on the mean of Y? To approach this question, imagine that the decision to administer treatment at each time t were made totally at random by the treating physician. In that hypothetical case, giving treatment at time t is not expected to be associated with any measured or unmeasured prognostic factors (i.e., there would be no "confounding") and therefore γ_2 would intuitively have a causal interpretation. Similarly, γ_2 would keep its causal interpretation if the physician's decision were based only on the history of treatment prior to t. Whenever the conditional probability of receiving treatment on day t given past treatment and prognostic factors history (measured and unmeasured) depends only on past treatment history, we say the process is a "causally exogenous or ancillary treatment process". (A more formal mathematical definition is provided below.) It is well-recognized in the social sciences, econometrics, epidemiologic, and biostatistical literature that γ_2 will have a causal interpretation if $A(t)$ is a causally exogenous (or ancillary) covariate process. Randomized treatments like the one described above are causally exogenous treatments.

We say that a treatment $A(t)$ is a "statistically exogenous or ancillary process" if the probability of receiving treatment at time t does not depend on the history of measured time-dependent prognostic factors $\overline{L}(t)$ up to t conditional on treatment history prior to t, i.e.,

$$\overline{L}(t)\, A(t) \mid \overline{A}(t-1),$$

where $A \coprod B \mid C$ *means that* A *is independent of* B *given* C.

Note that a necessary condition for $A(t)$ to be "causally exogenous" is for it to be "statistically exogenous." However, that a process is "statistically exogenous" does not imply it is "causally exogenous," because there may be unmeasured prognostic factors (i.e., confounders) that predict the probability of treatment $A(t)$ at time t given past treatment history. We can test from the data whether $A(t)$ is statistically exogenous but are unable to test whether a statistically exogenous process is causally exogenous.

Suppose $A(t)$ is discrete and we can correctly model the probability
$f\left[a(t) \mid \overline{\ell}(t), \overline{a}(t-1)\right]$ of receiving treatment $a(t)$ on day t as a function of past treatment $\overline{a}(t-1)$ and measured prognostic factor history $\overline{\ell}(t)$. We could then measure the degree to which the treatment process is statistically non-exogenous through day t by the random quantity

$$\mathcal{W}(t) = \prod_{k=0}^{t} f\left[A(k) \mid \overline{A}(k-1), \overline{L}(k)\right] / f\left[A(k) \mid \overline{A}(k-1)\right] \ .$$

The numerator in each term in $\mathcal{W}(t)$ is the probability that a subject received his own observed treatment at time k, $A(k)$, given his past treatment and prognostic factor history. The denominator is the probability that a subject received his observed treatment conditional on his past treatment history but not further adjusting for his past prognostic factor history. Note that the treatment process is statistically exogenous just in the case that $\mathcal{W}(t) = 1$ for all t. Of course, $\mathcal{W}(t)$ is unknown and will have to be estimated from the data but, for pedagogic purposes, assume for the moment that it were known.

When $A(t)$ is a statistically endogenous process, we shall consider estimating γ by a weighted logistic regression in which a subject is given the weight $\mathcal{W}^{-1} \equiv [\mathcal{W}(K)]^{-1}$. The weighted logistic regression estimator maximizes ${}_{i=1}^{n} \left[Lik_i(\gamma)\right]^{\mathcal{W}_i^{-1}}$. This weighted logistic regression would agree with the usual unweighted analysis described above just in the case in which $A(t)$ were exogenous. The somewhat surprising result described in detail below is that, if the vector of prognostic factors recorded in $L(t)$ constitutes all relevant time-dependent prognostic factors (i.e., confounders), then, whether or not the treatment process is statistically exogenous, the weighted logistic regression estimator of γ_2 will converge to a quantity β_2 that can be appropriately interpreted as the causal effect of treatment history on the mean of Y. In contrast, when $A(t)$ is statistically endogenous, the usual logistic regression estimator will still converge to the parameter γ_2, but now γ_2 will have no causal interpretation.

To prove such a claim, we need to give a formal mathematical meaning to the informal concept of the causal effect of treatment history on the mean of Y. To do so, we first introduce some notational conventions. We use capital letters to represent random variables and lower case letters to represent possible realizations (values) of random variables. For example, O_i is the random observed data for the i^{th} study subject and o is a possible realization (value) of O_i. Further, we assume that the random vector O_i for each subject is drawn independently from a distribution common to all subjects. Because the O_i have the same distribution, we often suppress the i subscript.

Counterfactual Outcomes:

Now we introduce counterfactual or potential outcomes. For any fixed non-random treatment history $\overline{a} =$
$\{a(u); 0 \leq u \leq K+1\}$, let $Y_{\overline{a}}$ be the random variable representing a subject's outcome had, possibly contrary to fact, the subject been treated with history $\overline{a}$ rather than his observed history $\overline{A}$. Note the $\overline{a}$'s are possible realizations of the random variable $\overline{A}$. For each possible history $\overline{a}$, we are assuming a

subject's response $Y_{\overline{a}}$ is well defined (although generally unobserved). Indeed we only observe $Y_{\overline{a}}$ for that treatment history $\overline{a}$ equal to a subject's actual treatment history $\overline{A}$, i.e., $Y = Y_{\overline{A}}$. Then formally our statement that the effect of treatment history on the mean of Y is a linear logistic function of cumulative treatment is the statement that, for each $\overline{a}$,

$$E\left[Y_{\overline{a}}\right] = g\left(\overline{a};\beta\right) \text{ } where \text{ } g\left(\overline{a};\beta\right) = \left[1+\exp\left\{-\beta_1-\beta_2 \text{ } cum\left(\overline{a}\right)\right\}\right]^{-1} \quad ,$$

which we refer to as a MSM for the effect of treatment on the mean of Y.The model for $E\left[Y_{\overline{a}}\right]$ is a marginal structural model since it is a model for the marginal distribution of counterfactual variables and, in the econometric and social science literature, causal models (i.e., models for counterfactual variables) are often referred to as structural.

The parameter β_2 of our MSM is of important policy interest. To see why, consider a new subject exchangeable with (i.e., drawn from the same distribution as) the n study subjects. We must decide which treatment history $\overline{a}$ to administer to the new subject. We would like to provide the treatment that minimizes the subject's probability of having HIV RNA in his blood at end of follow-up. That is, we want to find $\overline{a}$ that minimizes $E\left[Y_{\overline{a}}\right]$. Thus, for example, if the parameter β_2 of our causal model is positive, we will withhold AZT treatment from our subject (i.e., we will give him the treatment history $\overline{a} \equiv 0$), since positive β_2 indicates that the probability of having HIV RNA in one's blood at the end of follow-up increases with increasing cumulative AZT dose. In contrast to β_2, the parameter γ_2 of our association model $E\left[Y \mid \overline{A}\right] = g\left(\overline{A};\gamma\right)$ may have no causal interpretation. For example, suppose physicians preferentially started AZT on subjects who, as indicated by their prognostic factor history, were doing poorly and that AZT has no causal effect on the mean of Y (i.e., $\beta_2 = 0$). Nonetheless, the mean of Y will increase with cumulative AZT doses and thus γ_2 will be positive. In this setting, we say that the parameter γ_2 of the association model lacks a causal interpretation because it is confounded by the association of the prognostic factors $\overline{L}\left(u\right)$ with the treatment $A\left(u\right)$.

Formally, in terms of counterfactuals, we say that the $A\left(t\right)$ process is "causally exogenous" if, for all histories $\overline{a}$,

$$Y_{\overline{a}} A\left(t\right) \mid \overline{A}\left(t-1\right)$$

which is equivalent to

$$Y_{\overline{a}} \coprod \overline{A} \text{ }.$$

Given the covariates recorded in $L\left(t\right)$, we say there are no unmeasured confounders if for each $\overline{a}$

$$Y_{\overline{a}} A\left(t\right) \mid \overline{L}\left(t\right), \overline{A}\left(t-1\right) \quad .$$

With these formalizations, it can then be shown mathematically, that when there are no unmeasured confounders, *(i)* statistical exogeneity [i.e., $\overline{L}\left(t\right) A\left(t\right) \mid \overline{A}\left(t-1\right)$] implies that the $A\left(t\right)$ process is "causally exogenous," *(ii)* the weighted logistic estimator converges to the parameter β_2 of the marginal structural model for $E\left[Y_{\overline{a}}\right]$, and *(iii)* the limit γ_2 of the usual logistic estimator generally differs from the causal parameter β_2 of the MSM unless the treatment process is statistically exogenous.

We shall also refer to the assumption of no unmeasured confounders as the assumption that treatment $A\left(t\right)$ is sequentially ignorable or randomized given the past. The assumption states that, conditional on AZT history and the history of all recorded covariates prior to t, increments in AZT dosage rate at t are independent of the counterfactual random variables $Y_{\overline{a}}$. This assumption will be true if all prognostic factors for, i.e., predictors of, $Y_{\overline{a}}$ that are used by patients and physicians to determine the dosage of

AZT at t are recorded in $\overline{L}(t)$ and $\overline{A}(t-1)$. For example, since physicians tend to withhold AZT from subjects with low white blood count, and in untreated subjects, low white blood count is a predictor of HIV RNA, the assumption of no unmeasured confounders would be false if $\overline{L}(t)$ does not contain WBC history. It is the primary goal of the epidemiologists conducting an observational study to collect data on a sufficient number of covariates to ensure that the assumption of no unmeasured confounders will be at least approximately true.

The assumption of no unmeasured confounders is the fundamental condition that will allow us to draw causal inferences from observational data. It is precisely because it cannot be guaranteed to hold in an observational study and is not empirically testable that it is so very hazardous to draw causal inferences from observational data. Note that if, as in a sequentially randomized trial, at each time t, the dose of AZT was chosen at random by the flip of a coin, then the assumption of no unmeasured confounders would be true even if the probability that the coin landed heads depended on past measured covariate and AZT-history. It is because physical randomization guarantees the assumption that most people accept that valid causal inferences can be obtained from a randomized trial. See Rubin (1978), Robins (1986) and Holland (1986) for further discussion. Robins (1997, 1998b) and Robins et al. (1999) discuss how the consequences of violations of the assumption of no unmeasured confounders can be explored through sensitivity analysis. Also see Appendix 3 below.

Given the assumption of no unmeasured confounders, Robins (1987) shows the mean of the dichotomous variable $Y_{\overline{a}}$ is non-parametrically identified from the joint distribution F_O of the observed data O by the g-computation algorithm formula of Robins (1986). Specifically, $E(Y_{\overline{a}}) = b(\overline{a})$ where

$$b(\overline{a}) \equiv \int \cdots \int E\left(Y \mid \overline{\ell}_K, \overline{a}_K\right)_{k=0}^{K} f\left(\ell_k \mid \overline{\ell}_{k-1}, \overline{a}_{k-1}\right) d\mu\left(\ell_k\right) \qquad (*)$$

and for notational convenience we have written $\overline{z}(k)$ as $\overline{z}_k$ and $z(k)$ as z_k.

The g-computation algorithm functional $b(\overline{a})$ is the marginal mean of Y in the manipulated subgraph of the directed acyclic graph (DAG) G representing the observed data O in which all arrows into the treatment variables $\overline{A} = (A_1, \ldots, \overline{A}_K)$ have been removed and $\overline{A}$ is set to $\overline{a}$ with probability 1 (Spirtes et al., 1993). More specifically, let DAG G be the complete DAG with temporally ordered vertex set $O = \{L_0, A_0, L_1, A_1, \ldots, A_K, Y\}$ and let DAG $G_{\overline{a}}$ be the subgraph of G in which all arrows into the $A_k, k = 0, \ldots, K$ have been cut. Then $b(\overline{a})$ is the marginal mean of Y based on a distribution for O represented by DAG $G_{\overline{a}}$ in which $f\left(A_k \mid \overline{A}_{k-1}, \overline{L}_k\right)$ is replaced by a degenerate density that takes the value a_k with probability 1 while the conditional density of each other variable in the set O given its parents remains as in F_O.

We say that the distribution of $O = \{L_0, A_0, L_1, A_1, \ldots, A_K, Y\}$ is standardly parameterized if, for each variable in O, we have specified a parametric or semiparametric model for the conditional distribution of that variable given its temporal predecessors (the past) and the parameters of each conditional model are variation-independent of those of any other conditional model. When our goal is to estimate the effect of a sequential (time-dependent) treatment $\overline{A}$ on an outcome Y, Lemma 1 and Theorem 2 of Robins and Wasserman (1997) imply that inference procedures based on the standard parameterization will fail. Specifically, they prove that common choices for the parametric families in a standard parameterization often lead to joint densities such that the g-computation formula for $E(Y_{\overline{a}})$ can never satisfy the causal null hypothesis that $E(Y_{\overline{a}})$ is the same for all $\overline{a}$. In particular, the causal null hypothesis does not imply that $Y \coprod \overline{A}_K \mid \overline{L}_K$. As a consequence, in large samples, the causal null hypothesis, even when true, will be falsely rejected regardless of the data. Robins and Wasserman propose reparameterizing the distribution of O using structural nested models. MSMs represent an alternative reparameterization that also overcomes the fatal deficiencies of the standard parameterization.

Theory of Inverse-Probability-of-Treatment-Weighting: We now explain why weighting by $\mathcal{W}^{-1}$ corrects our logistic regression estimator for the "confounding" due to the prognostic factors in $L(t)$. The first point to note is that in the definition of $\mathcal{W}(t)$ we could have replaced the denominator $pr\left[A(t) \mid \overline{A}(t-1)\right]$ by any other function of $\overline{A}(t)$ without influencing the consistency of our weighted logistic estimator of the parameter β_2 of the MSM; only the efficiency (variance) of our estimator would be influenced. However, our estimator would be inconsistent if we replaced the numerator by any other function of $\overline{A}(t)$ and $\overline{L}(t)$. Thus one can view weighting by $\mathcal{W}^{-1}$ as weighting by the inverse of a subject's probability of having his own observed treatment history. Now view each person as a member of a pseudo- or ghost population consisting of themselves and $\mathcal{W}^{-1}-1$ ghosts (copies) of themselves who have been added by weighting. In this new ghost or pseudo population, it is easy to show that $\overline{L}(t)$ does not predict treatment at t given past treatment history, and thus we have created a pseudo-population in which treatment is exogenous. Furthermore, the causal effect of $\overline{A}$ on Y in the ghost population is the same as in the original population. That is, if $E\left[Y_{\overline{a}}\right]=g(\overline{a};\beta)$ in the true population, the same will be true of the ghost population. Hence, we would like to do ordinary logistic regression in the pseudo-population. That is essentially what our weighted logistic regression estimator is doing, since the weights create, as required, $\mathcal{W}^{-1}-1$ additional copies of each subject.

A formal, mathematical explanation of why weighting by W^{-1} corrects our logistic regression estimator for "confounding" is given in the following lemma characterizing the g-computation algorithm functional $b(\overline{a})$ defined in (*) above.

Lemma 1.1: *$b(\overline{a})$ defined in (*) is the unique function $c(\overline{a})$ of $\overline{a}$ such that $E\left[q\left(\overline{A}\right)\left(Y-c\left(\overline{A}\right)\right)/\mathcal{W}\right]=0$ for all functions $q\left(\overline{A}\right)$ for which the expectation exists.*

Lemma 1.1 has the following corollary.

Lemma 1.2: *Under sequential randomization, $E\left(Y_{\overline{a}}\right)$ is unique function $c(\overline{a})$ of $\overline{a}$ such that $E\left[q\left(\overline{A}\right)\left(Y-c\left(\overline{A}\right)\right)/\mathcal{W}\right]=0$ for all functions $q\left(\overline{A}\right)$ where the expectation exists.*

Consistency of our weighted estimator then follows from the fact that the probability limit of our weighted score equation is $E\left[q\left(\overline{A}\right)\left(Y-c\left(\overline{A}\right)\right)/\mathcal{W}\right]=0$ with $q\left(\overline{A}\right)=(1,cum(\overline{A}))'$ and $c\left(\overline{A}\right)=g(\overline{A},\beta)$.

Under a mild strengthening of our assumption of sequential randomization (no unmeasured confounders), a simple, quite revealing, purely "causal" proof of Lemma 1.2 can be obtained that does not use the fact that $E\left(Y_{\overline{a}}\right)$ is given by the g-computation algorithm formula $b(\overline{a})$ of Eq. (*). Let $Y_{\overline{\mathcal{A}}}=\left\{Y_{\overline{a}};\overline{a}\in\overline{\mathcal{A}}\right\}$ where $\overline{\mathcal{A}}$ is the support of the random variable $\overline{A}$. Suppose we strengthen our assumption of no unmeasured confounders to

$$Y_{\overline{\mathcal{A}}}\coprod A_k \mid \overline{L}_k,\overline{A}_{k-1}\ .$$

Denote the factual and counterfactual data by $Z=\left(Y_{\overline{\mathcal{A}}},\overline{A},\overline{L}\right)$ and the observed data by $O=\left(Y\equiv Y_{\overline{A}},\overline{A},\overline{L}\right)$. We can factor the true joint density of Z that generated the data as

$$f(Z)=f\left(Y_{\overline{\mathcal{A}}}\right)\prod_{k=0}^{K}f\left(L_k\mid\overline{L}_{k-1},\overline{A}_{k-1},Y_{\overline{\mathcal{A}}}\right)\prod_{k=0}^{K}f\left(A_k\mid\overline{L}_k,\overline{A}_{k-1}\right)\ .$$

Now let $f^*\left(A_k\mid\overline{A}_{k-1}\right)$ be a density for A_k given $\overline{A}_{k-1}$. It need not equal the true density $f\left(A_k\mid\overline{A}_{k-1}\right)$. Let $f^*(Z)$ be a joint density for Z that differs from the true joint density $f(Z)$ only in that $f^*\left(A_k\mid\overline{L}_k,\overline{A}_{k-1}\right)=f^*\left(A_k\mid\overline{A}_{k-1}\right)$ so that A_k is strictly exogenous. Thus,

$$f^*(Z)=f\left(Y_{\overline{\mathcal{A}}}\right)\prod_{k=0}^{K}f\left(L_k\mid\overline{L}_{k-1},\overline{A}_{k-1},Y_{\overline{\mathcal{A}}}\right)\prod_{k=0}^{K}f^*\left(A_k\mid\overline{A}_{k-1}\right)\ .$$

Now $E(Y_{\overline{a}}) = E^*(Y_{\overline{a}})$ since $f(Z)$ and $f^*(Z)$ have the same marginal law for $Y_{\overline{a}}$. Second, since $\overline{A}$ is causally exogenous under $f^*(z)$ [i.e., $Y_{\overline{a}}^* \overline{A}$], we have that $E^*[Y_{\overline{a}}] = E^*[Y_{\overline{a}} \mid \overline{A} = \overline{a}] = E^*[Y_{\overline{A}} \mid \overline{A} = \overline{a}] = E^*[Y \mid \overline{A} = \overline{a}]$. That is, by $\overline{A}$ causally exogenous, the mean of $Y_{\overline{a}}$ is given by the regression function $E^*[Y \mid \overline{A} = \overline{a}]$ of Y on $\overline{A} = \overline{a}$. Now it is a standard result that the regression function $E^*(Y \mid \overline{A} = \overline{a})$ is characterized as the unique function $c(\overline{a})$ solving $E^*\{q(\overline{A})[y - c(\overline{A})]\} \equiv \int q(\overline{A})(y - b(\overline{A})) f^*(Z)\, d\mu(Z) = 0$ for all $q(\overline{A})$ where μ is a dominating measure. But, $\int q(A)(Y - c(\overline{A})) f^*(Z)\, d\mu(Z) = \int q(A)(Y - c(\overline{A})) \frac{f^*(Z)}{f(Z)} f(Z) d\mu(Z) = E\left[q(A)(Y - c(\overline{A})) \frac{f^*(Z)}{f(Z)}\right]$. But, by definition, $\frac{f^*(Z)}{f(Z)} = \mathcal{W}^{-1}$ when $f^*(A_k \mid \overline{A}_{k-1}) = f(A_k \mid \overline{A}_{k-1})$. Lemma 1.2 then follows, since $E^*(Y \mid A = \overline{a}) = E(Y_{\overline{a}})$. The proof also makes clear that consistency of our weighted estimator does not require that we choose $f^*(A_k \mid \overline{A}_{k-1}) = f(A_k \mid \overline{A}_{k-1})$.

Data Analyses: Marginal Structural Mean Model for a Repeated Measures Outcome:

To give a better picture of the meaning and use of MSMs, we report preliminary results of two data analyses. Full details will be published elsewhere. We estimate the joint effect in ACTG Randomized Trial 002 of AZT treatment arm and aerosolized pentamidine (AP) on the evolution of CD4 count in the first analysis and on mortality in the second analysis. This trial was designed to compare the effect of high-dose AZT with low-dose AZT on survival. However, over fifty percent of the subjects failed to comply with the assigned treatment protocol and initiated treatment with a non-randomized therapy, AP, during the course of the trial. The joint effects of AP and AZT treatment arm on survival have been previously estimated using structural nested failure time models by Robins and Greenland (1994). We first consider a MSM model for the effect of AP therapy on the mean of the log transformed CD4 count history while adjusting for baseline variables. CD4 count measurements were obtained at weeks 8, 16, 24, and 32 measured in days. Specifically, we consider the MSM

$$E\left[Y_{\overline{a}}(m) \mid V^{\dagger}\right] = g_m\left[\overline{a}(m), V^{\dagger}, \beta\right]$$

where $Y(m) = \log[CD4(m) + 2]$, $CD4(m)$ is the CD4 count on day m, $Y_{\overline{a}}(m)$ is the counterfactual version of $Y(m)$ under the AP history $\overline{a}$, $V^{\dagger} = (1, m, R, Y(0), \log WBC(0))'$ is the vector of baseline regressors with $R = 1$ denoting the high AZT treatment arm and $R = 0$ the low AZT treatment arm, $Y(0)$ is defined above, and $WBC(0)$ is baseline white blood count. We modeled the regression function as $g_m\left[\overline{a}(m), V^{\dagger}, \beta\right] = \beta_1' V^{\dagger} + \beta_2 cum(m, \overline{a})$ with $cum(m, \overline{a}) =_0^m a(t)\, dt$ being cumulative AP treatment up to day m. We make the assumption of no unmeasured confounders with $L(t)$ being white blood count, the number of episodes of pneumocystis pneumonia (PCP), an AIDS-related pneumonia, up to day t, and CD4 count on day t. Furthermore, the baseline covariates $V^{\dagger}$ are included in $L(0)$. Arguing as above, it can be shown a consistent estimator of β is obtained by fitting the model $E\left[Y(m) \mid V^{\dagger}, \overline{A}\right] = g_m\left[\overline{A}(m), V^{\dagger}, \beta\right]$ using the generalized estimating equation (GEE) option of Proc genmode in the SAS software package under a working independence matrix and weighting the observation $Y(m)$ for a subject by $\{\mathcal{W}(m)\}^{-1}$. The GEE option of Proc genmode is simply a program that fits the above model by weighted least squares to obtain an estimate of β. In estimating β, the program treats each individual at each of the four times m as four separate observations when computing the least squares estimator. However, the program outputs a robust variance estimator that appropriately accounts for the fact that the four observations on a given subject are correlated.

Since $\mathcal{W}(m)$ was unknown, it was estimated from the data by fitting logistic models for $pr\left[A(k) \mid \overline{L}(k), \overline{A}(k-1)\right]$ and $pr\left[A(k) \mid \overline{A}(k-1), V^{\dagger}\right]$. [Note that when our MSM conditions on baseline variables $V^{\dagger}$, they should be included in the denominator of $\mathcal{W}(m)$.] Specifically, we fit the model

$$logit\, pr\left[A(k) = 0 \mid \overline{L}(k), \overline{A}(k-1) \equiv 0\right] = \alpha' Q(k)$$

where $Q(k) = (1, \log k, \log[WBC(k-1)], Y(k-1), PCP\,bouts(k-1), Y(0), \log[WBC(0)], R)$. Here, PCP bouts $(k-1)$ is the number of episodes (bouts) of pneumocystis pneumonia through $k-1$. In fitting the model, we treated each subject at each day k, $k = 0, 1, \ldots, 224$ as an independent observation (which is justified by the conditional martingale structure of the model). We note that since in the 002 data file, any subject starting AP remained on it thereafter, it was only necessary to fit a model for $A(k)$ for subjects who had yet to begin AP [i.e., $\overline{A}(k-1) \equiv 0$]. We estimated $pr\left[A(k) \mid \overline{A}(k-1) \equiv 0, V^{\dagger}\right]$ by fitting the above model after eliminating the random time-dependent terms that were functions of $(k-1)$. The 95 percent Wald intervals computed using the robust variance outputted by the GEE program are conservative (i.e., they are guaranteed to cover the true β at least 95 percent of the time in large samples) because they do not account for estimation of the weights $\mathcal{W}(m)$. It is interesting that estimating the weights shrinks the variance of our estimator of β, so that our intervals (which do not account for the fact that the weights are estimated) are conservative. Note that the elements of the vectors α multiplying the time-dependent covariates $\log[WBC(k-1)]$, $Y(k-1)$ and PCP bouts $(k-1)$ will all be zero if and only if the AP treatment process is statistically ancillary. A three degree of freedom likelihood ratio test of the hypothesis that all three components of α were zero rejected at the $p < .01$ level. As a consequence, we rejected the hypothesis of statistical exogeneity.

The analysis just described assumes that there is no drop-out or censoring by end of follow-up. To correct for this, we defined a subject as censored (i.e., permanently missing) the first time he missed one of his scheduled visits or was censored by end of follow-up. Under the assumption of ignorable drop-out given the time-dependent factors $L(t)$ and treatment $A(t)$, we still obtain consistent estimators of β in the presence of drop-out if we weight a subject uncensored at day m by $\left\{\mathcal{W}(m)\,\mathcal{W}^{\dagger}(m)\right\}^{-1}$ where $\mathcal{W}^{\dagger}(m) =_{k=0}^{m}$ $\{p\left[R(k)=0 \mid \overline{R}(k-1)=0, \overline{L}(k-1), \overline{A}(k-1)\right] / pr\left[R(k)=0 \mid \overline{R}(k-1)=0, \overline{A}(k-1), V^{\dagger}\right]\}$ is the ratio of a subject's probability of remaining uncensored up to day m divided by that probability calculated as if there had been no time-dependent determinants of drop-out except past treatment history. Here, $R(m) = 0$ if a subject has not dropped out or reached end to follow-up by day m. Since $\mathcal{W}^{\dagger}(m)$ is unknown, it was estimated from the data in a manner completely analogous to the estimation of $\mathcal{W}(m)$ except with $A(k)$ replaced by $R(k)$ as the outcome variable and with $A(k-1)$ added as an additional regressor. Furthermore, in the presence of censoring, it is necessary when estimating $pr[A(k)=0 \mid \cdot]$ to add the event $R(k)=0$ to the conditioning event.

We fit the above models and obtained an estimate $\widehat{\beta}_2 = .001$ and conservative 95 percent confidence interval $(-.026, .028)$ for the parameter β_2 representing the causal effect of cumulative AP dose on CD4 count. Furthermore, since in trial 002 the assignment to AZT treatment arm was at random with probability 1/2, the AZT treatment arm indicator is exogenous. It follows that the component β_{1R} of β_1 multiplying the AZT treatment indicator R has the interpretation as the direct effect of AZT treatment arm on the evolution of mean CD4 count that is not through AP history. We obtained an estimate of .0196 with a conservative 95 percent confidence interval of $(-.123, .163)$ for β_{1R}.

Marginal Structural Cox Proportional Hazards Model: We next estimated the joint effects of AP therapy and AZT treatment arm on survival by specifying a marginal structural Cox proportional hazards model

$$\lambda_{T_{\overline{a}}}\left(t \mid V^{\dagger}\right) = \lambda_0(t)\exp\left[\beta_1' V^{\dagger} + \beta_2 a(t)\right]$$

where $T_{\overline{a}}$ is the subject's time to death if he had followed AP history $\overline{a}$, $\lambda_{T_{\overline{a}}}\left(t \mid V^{\dagger}\right)$ is the hazard of $T_{\overline{a}}$ at t given $V^{\dagger}$, $\lambda_0(t)$ is an unspecified baseline hazard function, and $V^{\dagger} = (R, Y(0), \log WBC(0))$. Note this model specifies that the hazard of failure at time t depends on current AP status rather than on cumulative AP history. Arguing as in the previous subsection, we can obtain consistent estimates of the unknown parameter $\beta = (\beta_1', \beta_2)'$ by fitting the ordinary time-dependent Cox model $\lambda_T\left(t \mid \overline{A}(t), V^{\dagger}\right) = \lambda_0(t)\exp\left[\beta_1' V^{\dagger} + \beta_2 A(t)\right]$ except that the contribution of subject to a calculation performed on subjects at risk at time t is weighted by $\widehat{\mathcal{W}}(t)^{-1}\widehat{\mathcal{W}}^{\dagger}(t)^{-1}$. Note that now when we model $pr[R_k = 0 \mid \cdot]$ and $pr[A_k = 0 \mid \cdot]$ we must include the event $T > k$ among the conditioning events. Here we have adopted the convention that on any day k censoring occurs at the end of the day. Note since the subject-specific weights change with time, one either needs to write a special program or trick a standard time-dependent Cox model that allows weights into allowing for time-varying weights by a clever use of the time-varying stratum option available in many off-the-shelf Cox programs. To obtain conservative 95 percent confidence intervals for β, one needs to compute the so-called robust variance of Lin et al. (1989). Implementing the above procedure, we obtained an estimate $\widehat{\beta}_2 = -.1362$ with 95 percent conservative confidence interval $(-.35, .09)$ for β_2. For the component β_{1R} of β_1 representing the direct effect of AZT treatment arm on survival, we obtained an estimate of .1890 with a 95 percent confidence interval of $(-.01,.21)$ indicating borderline statistically significant evidence for a beneficial effect of the low-dose AZT arm. Both the results obtained for the prophylaxis effect and for the AZT effect were consistent with those obtained by Robins and Greenland (1994) using SNFTMs.

Philosophical Interlude: We pause to comment briefly on the definition and nature of the counterfactual random variables $T_{\overline{a}}$. Following Lewis (1973), we consider $T_{\overline{a}}$ to be a subject's death time in the closest possible world to this in which, possibly contrary to fact, the subject was treated with the AP history $\overline{a}$. Consider a subject i who, in the 002 trial, was assigned to the high-dose AZT arm, received AP from week 10 to 40, took the assigned 1500 mg. of AZT daily until week 12 but then stopped all further AZT therapy, and finally died in week 40. If AP had been withheld, it is quite conceivable that subject i would have continued to be assigned 1500 mg. of AZT daily past week 12 if either *(1)* AP potentiated the toxic effects of AZT, precipitating a life-threatening toxic episode in week 12, or *(2)* the subject, although not toxic, had stopped AZT at week 12 because he felt himself to be adequately protected by the AP treatment. To be concrete, say, in the closest possible world, subject i would have continued to take 1500 mg. of AZT daily through week 14 and none thereafter if AP had been withheld. We now consider the meaning of the counterfactual $T_0 \equiv T_{\overline{a} \equiv 0}$ in which AP was always withheld. Then, by its definition, T_0 would be equal to subject i's failure time when the subject was assigned to the high-dose arm, never received AP, and took AZT daily through week 14 (rather than through week 12). Thus, T_0 might well differ from the counterfactual variable, say T_0^*, representing a subject's survival time in the closest possible world in which AP was withheld but, as in this world, AZT was stopped after week 12. Several comments are in order.

First, T_{0i} is conceptually rather well-defined, even if we do not observe what the particular subject i would have done about his AZT dose after week 10 in the absence of AP therapy, as T_{0i} is just subject i's outcome in the closest possible world to this one in which all AP therapy is withheld and all consequences which flow from that (including possibly taking AZT in week 12-14) are all allowed to occur. Second, it may be quite reasonable to make (at least to a good approximation) the assumption of no unmeasured confounders for T_0, in which case its distribution is non-parametrically identified by inverse-probability-of-treatment-weighting, (equivalently, by the g-computation algorithm formula) from the distribution of the observed data. The intuitive reason for this successful identification is that for a subset of the population (i.e., those who never did take AP), we do observe T_0, and under the assumption of no unmeasured confounders, we can appropriately reweight them by $\mathcal{W}^{-1}$ to construct a ghost population whose distribution of $T_0 = T$ is the same as that of T_0 in the true study population. Third, from a public health point of view, it is much more important to identify the distribution of T_0 than T_0^* since it is the distribution of T_0 that would result if we made the public policy decision to withhold AP therapy. Fourth, T_0^* may be more relevant than T_0 in a legal case against the manufacturers of AP. For example, the manufacturers would argue that they should not be held responsible for any damages if a subject's observed death time T equalled their counterfactual death time T_0^* (even if T differed from T_0 due to differences in the amount of AZT taken). Fifth, the distribution of T_0^* is not identified even in an experiment in which both AP and AZT are randomly assigned. Thus, no amount of data evidence will ever determine the distribution of T_0^*, even in a randomized experiment (Robins and Greenland, 1989).

Comparison with SNMs: Marginal structural models are an alternative to structural nested models. A SNM is model for the magnitude of the causal effect of a final brief blip of a time-dependent treatment at time t as a function of past time-dependent treatment and prognostic factor history. The causal parameter of a structural nested model is identified under the assumption of no unmeasured confounders. The essential difference between MSMs and SNMs is that SNMs model the magnitude of the effect of a treatment given at t as a function of the prognostic factor history up to t. In contrast, MSMs model the causal effect of treatment given at t only as a function of baseline prognostic factors. Sec. 5 below is devoted to describing what is known about the advantages and disadvantages of MSMs versus SNMs. Some of the advantages of MSMs were discussed above. Possible disadvantages include the following. *(i)* Inability to easily estimate the effects of dynamic treatment regimes (i.e., treatment plans where a subject's covariate history up to time k determines the treatment to be taken at k). Actual medical treatments are usually dynamic, since if a subject becomes toxic to a drug, the drug must be stopped. *(ii)* The inability to directly test the null hypothesis of no effect of any treatment regime (dynamic or non-dynamic) on outcome. *(iii)* The difficulty in performing likelihood-based inference for MSMs, since the likelihood is a computational nightmare. *(iv)* Lack of identifiability of the MSM model parameters when sequential ignorability holds for a so-called "instrumental variable" but not for the actual treatment of interest. *(v)* MSMs, in contrast to SNMs, cannot be used if there exists a value of ℓ_k, say $\ell_k = 0$, such that for all but one value of a_k, $f\left[a_k \mid \overline{\ell}_{k-1}, \ell_k = 0, \overline{a}_{k-1}\right] = 0$. An example would be a study of the effect of an occupational exposure on mortality with $\ell_k = 0$ if a subject is off work at time k, $\ell_k = 1$ otherwise, and subjects off work can only receive exposure level $a_k = 0$. We note that SNMs do not suffer from any of these five deficiencies.

2 Advantages of MSMs with Continuous Y or with Failure Time: A Formal Definition of MSMs

2.1 The Data

Consider a study where we observe n i.i.d. copies of data $O = \left(\overline{A}(C), \overline{L}(C)\right)$, where C is an administrative end of follow-up time, $\overline{A}(C)$ is a treatment process, $\overline{L}(C)$ is an outcome or response process and, for any $Z(u), \overline{Z}(t) \equiv \{Z(u); 0 \le u \le t\}$. We assume C is an element of $L(0)$ since it is assumed known at time 0.

For purposes of causal inference, we assume the existence of an underlying treatment process $\overline{A} = \{A(u); 0 \le u < \infty\}$ with $A(u)$ taking values in a set $\mathcal{A}(u)$ and the existence of underlying counterfactual random variables

$$\{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\} \tag{1}$$

where $\overline{L}_{\overline{a}} = \{L_{\overline{a}}(u)\,; 0 \leq u < \infty\}$, $\overline{a} = a(\cdot) = \{a(t)\,; 0 \leq t < \infty$ and $a(t) \in \mathcal{A}(t)\}$ is a treatment plan (equivalently, regime or function) lying in a set of functions $\overline{\mathcal{A}}$. Given a regime $\overline{a}$, let $\overline{L}_{\overline{a}(u),\mathbf{0}}$ be counterfactual history under a regime $\overline{a}^*$ that agrees with $\overline{a}$ through time u and is 0 thereafter, where 0 is the baseline value of $a(t)$. Then we assume that the $\overline{L}_{\overline{a}}$ satisfy the following consistency assumption with probability 1:

$$\overline{L}_{\overline{a}(u),\mathbf{0}}(u) = \overline{L}_{\overline{a}(t),\mathbf{0}}(u) = \overline{L}_{\overline{a}}(u) = \overline{L}_{\overline{a}^\dagger}(u) \tag{2}$$

for all $t > u$ and all $\overline{a}^\dagger$ with $\overline{a}^\dagger(u) = \overline{a}(u)$. This assumption essentially says that the future does not determine the past. The observed data are linked to the counterfactual data by

$$\overline{L}(C) = \overline{L}_{\overline{A}(C),0}(C) \ . \tag{3}$$

Eq. (3) states that a subject's observed outcome history through end of follow-up is equal to their counterfactual outcome history corresponding to the treatment they did indeed receive. We assume $\overline{L}_{\overline{a}} = \left(\overline{Y}_{\overline{a}}, \overline{V}_{\overline{a}}\right)$ where $\overline{Y}_{\overline{a}}$ is an outcome process of interest and $\overline{V}_{\overline{a}}$ is the process of other recorded variables. Further, we shall make the sequential randomization (i.e., ignorable treatment assignment) assumption that for all t and $\overline{a} \in \overline{\mathcal{A}}$,

$$\underline{Y}_{\overline{a}}(t) \coprod A(t) \mid \overline{L}\left(t^-\right), \overline{A}\left(t^-\right) \tag{4}$$

where for any variable $\underline{Z}(t) = \{Z(u)\,; u \geq t\}$ is the history of that variable from t onwards. We also refer to (4) as the assumption of no unmeasured confounders given prognostic factors $L(t)$. Because of measurability issues, (4) is not well-defined. If the $A(t)$ process can only jump at discrete non-random times $t_1, t_2, \ldots$ and the $\overline{L}(t)$ process has left-hand limits, i.e., $\overline{L}(t^-) \equiv \lim_{u \uparrow t} \overline{L}(u)$, (4) is formally, for each t_k,

$$f\left[A(t_k) \mid \overline{L}\left(t_k^-\right), \overline{A}\left(t_k^-\right), \underline{Y}_{\overline{a}}(t_k)\right] = f\left[A(t_k) \mid \overline{L}\left(t_k^-\right), \overline{A}\left(t_k^-\right)\right] \ . \tag{5}$$

where $f(\cdot \mid \cdot)$ is the conditional density of $A(t_k)$ with respect to a dominating measure. If $A(t)$ is a marked point process that can jump in continuous time with CADLAG (continuous from the right with left-hand limits) step-function sample paths, then Eq. (4) is formally that

$$\lambda_A\left[t \mid \overline{L}\left(t^-\right), \overline{A}\left(t^-\right), \underline{Y}_{\overline{a}}(t)\right] = \lambda_A\left[t \mid \overline{L}\left(t^-\right), \overline{A}\left(t^-\right)\right] \tag{6a}$$

and

$$\begin{aligned} &f[A(t) \mid \overline{L}\left(t^-\right), \overline{A}\left(t^-\right), A(t) \neq A\left(t^-\right), \underline{Y}_{\overline{a}}(t)] = \\ &\quad f\left[A(t) \mid \overline{L}\left(t^-\right), \overline{A}\left(t^-\right), A(t) \neq A\left(t^-\right)\right] \ . \end{aligned} \tag{6b}$$

Here, the intensity process $\lambda_A(t \mid \cdot)$ is $\lim_{\delta t \to 0} pr[A(t+\delta t) \neq A(t^-) \mid A(t^-), \cdot]/\delta t$. Eq. (6a) says that given past treatment and confounder history, the probability that the A process jumps at t does not depend on the future counterfactual history of the outcome of interest. Eq. (6b) says that given that the covariate process did jump at t, the probability it jumped to a particular value of $A(t)$ does not depend on the future counterfactual history of the outcome of interest.

Following Heitjan and Rubin (1991), we say the data are coarsened at random (CAR) if

$$f\left[\overline{A}(C) \mid \left\{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\right\}\right] \ \textit{depends only on}\ O = \left(\overline{A}(C), \overline{L}(C)\right) \ . \tag{7}$$

Note that we can use ideas from the "missing data" literature because one's treatment history $\overline{A}(C)$ determines which components of one's counterfactual history $\left\{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\right\}$ one observes. Thus we can view causal inference as a missing data problem (Rubin, 1976). If, as in all the models we shall consider in this paper, for each $\overline{\mathcal{A}}^{\dagger} \subseteq \overline{\mathcal{A}}$ satisfying $\overline{a}_1(u) \neq \overline{a}_2(u)$ for all $\overline{a}_1, \overline{a}_2 \in \overline{\mathcal{A}}^{\dagger}$, the $\left\{L_{\overline{a}}(u); \overline{a} \in \overline{\mathcal{A}}^{\dagger}\right\}$ may have a non-degenerate joint distribution, then CAR implies sequential randomization (4) but the converse is not true (Robins et al., 1999). Robins (1997, pg. 83) gives examples where one would expect (4) to be true even when (7) is false. In this paper, we shall only need (4). However, even if (7) is also imposed this, by itself, essentially places no restrictions on the joint distribution of the observable random variables (Gill, van der Laan, Robins, 1997) and, thus, is not subject to empirical test.

2.2 MSMs

A MSM for $\{\overline{Y}_{\bar{a}}; \overline{a} \in \overline{\mathcal{A}}\}$ places restrictions on the marginal distribution of the $\overline{Y}_{\overline{a}}$ possibly conditional on a baseline variable $V^{\dagger}$ in $V(0)$ (with $C \in V^{\dagger}$ if C is random). Examples of MSMs follow. Each of these examples will be important in our comparison of MSMs with structural nested models below.

Model 1: Suppose $C = K+1$ w.p.1., the $\overline{A}(C)$ process jumps only at times $0, 1, 2, \ldots, K$ and the $\overline{L_{\overline{a}}}$ process jumps only at times $0^-, 1^-, 2^-, \ldots, K^-, K+1^-$. In models 1a-1c, we are only concerned with an outcome measured at end of follow-up. Hence, we set $Y_{\overline{a}}(m) \equiv 0$ with probability 1 for $m \leq K$ and define $Y_{\overline{a}} = Y_{\overline{a}}(K+1)$. Then we have

Model 1a – non-linear least squares: $E\left[Y_{\overline{a}} \mid V^{\dagger}\right] = g\left[\overline{a}(K), V^{\dagger}, \beta_0\right]$ where $g(\cdot,\cdot,\cdot)$ is a known function. This is the logistic regression MSM we discussed in the Introduction.

Model 1b – semiparametric regression: $\eta\left\{E\left[Y_{\overline{a}} \mid V^{\dagger}\right]\right\} = g\left[\overline{a}(K), V^{\dagger}, \beta_0\right] + g^{\dagger}\left(V^{\dagger}\right)$ where $\eta(\cdot)$ is a known monotone link function, $g^{\dagger}(\cdot)$ is unknown and unrestricted and $g(\cdot,\cdot,\cdot)$ is a known function satisfying $g\left(\mathbf{0}, V^{\dagger}, \beta\right) = 0$. The requirement that $g\left(\mathbf{0}, V^{\dagger}, \beta\right) = 0$ implies that $g^{\dagger}\left(V^{\dagger}\right)$ is the "main effect of $V^{\dagger}$." Such models are also referred to as partial spline models. They are semiparametric because the main effect of $V^{\dagger}$ is modelled non-parametrically.

Model 1c – stratified transformation model: $pr\left[R(\overline{a}, \beta_0) < t \mid V^{\dagger}\right] = F_0\left(t \mid V^{\dagger}\right), F_0\left(t \mid V^{\dagger}\right)$ an unknown distribution function, $R(\overline{a}, \beta) = r\left(Y_{\overline{a}}, \overline{a}, V^{\dagger}, \beta\right)$ is a known increasing function of $Y_{\overline{a}}$ satisfying $r\left(y, \overline{a}, V^{\dagger}, \beta\right) = y$ if $\overline{a} \equiv \mathbf{0}$ or $\beta = 0$. This model says that we know the conditional quantile-quantile function linking the $Y_{\overline{a}}$'s given $V^{\dagger}$ up to an unknown parameter β. It is the natural generalization of model 1b for mean functions to quantile-quantile functions.

In the following model, we are interested in the outcome at each $m \geq 1$ so we no longer assume that $Y_{\overline{a}}(m) \equiv 0$ with probability 1.

Model 1d – multivariate non-linear least squares: $E\left[Y_{\overline{a}}(m) \mid V^{\dagger}\right] = g_m\left[\overline{a}(m-1), V^{\dagger}, \beta_0\right]$, $m = 1, \ldots, K+1$ where the $g_m(\cdot,\cdot,\cdot)$ are known. This is the natural MSM version of longitudinal generalized estimating equation models for marginal means (Liang and Zeger, 1986). It is the model we use to analyze the 002 CD4 count data in the Introduction.

Model 2: $C = \infty$, $Y_{\overline{a}}$ is a failure time process, i.e., $Y_{\overline{a}}$ jumps from 0 to 1 at some particular time and stays at 1. Then define the failure time $T_{\overline{a}}$ by the equation $Y_{\overline{a}}(T_{\overline{a}}) = 1$ and $Y_{\overline{a}}\left(T_{\overline{a}}^-\right) = 0$. Let $\lambda_0(t)$ and $\lambda_0\left(t \mid V^{\dagger}\right)$ be unknown non-negative functions of t and $\left(t, V^{\dagger}\right)$ respectively and, for any $Z, \lambda_Z(u)$ is the hazard of Z.

Model 2a – Cox proportional hazards model: $\lambda_{T_{\overline{a}}}\left[t \mid V^{\dagger}\right] = \lambda_0(t) \exp\left[r\left(\overline{a}(t^-), t, V^{\dagger}; \beta_0\right)\right]$ where $r()$ is a known function satisfying $r(\mathbf{0}, t, 0; \beta) = 0$. This is the model we use to analyze the 002 mortality data in the Introduction.

Model 2b **– stratified Cox proportional hazards model:** $\lambda_{T_{\overline{a}}}\left(t \mid V^{\dagger}\right) =$
$\lambda_0\left(t \mid V^{\dagger}\right) \exp\left[r\left(\overline{a}(t^-) t, V^{\dagger}; \beta_0\right)\right]$ where, now, $r\left(\mathbf{0}, t, V^{\dagger}; \beta\right) = 0$.

Model 2c **– stratified time-dependent accelerated failure time model:** $pr[r\left(T_{\overline{a}}, \overline{a}, V^{\dagger}, \beta_0\right) <$

$t \mid V^{\dagger}] =$
$F_0\left(t \mid V^{\dagger}\right)$ where $r\left(u, \overline{a}, V^{\dagger}, \beta\right) = r\left(u, \overline{a}\left(u\right), V^{\dagger}, \beta\right)$ is a known function increasing in its first argument satisfying $r\left(u, \mathbf{0}, V^{\dagger}, \beta\right) = u$. This model can also be written as

$$\lambda_{R\left(\overline{a}, \beta_0\right)}\left(t \mid V^{\dagger}\right) = \lambda_0\left(t \mid V^{\dagger}\right)$$

for $\overline{a} \in \overline{\mathcal{A}}$, where $R\left(\overline{a}, \beta\right) = r\left(T_{\overline{a}}, \overline{a}, V^{\dagger}, \beta\right)$. This is the extension of model 1c to a failure time variable. It is the model studied by Robins and Tsiatis (1992).

3 Estimation

3.1 Ancillary treatment process

In this section, we consider estimation of the parameter β_0 of our marginal structural models. In this subsection, we will suppose that $\overline{A}$ is a causally ancillary (i.e., exogenous) covariate process, i.e.,

$$\overline{A} \coprod \left\{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\right\} | V^{\dagger} \ . \tag{8}$$

The often unrealistic assumption (8) implies CAR but, in contrast to CAR, places restrictions on the joint distribution of the data. Specifically (8) implies statistical ancillarity

$$A\left(t\right) \coprod \overline{L}\left(t^{-}\right) \mid \overline{A}\left(t^{-}\right), V^{\dagger} \tag{9}$$

and thus (8) is subject to an empirical test.

Given (8), the restrictions on the observables O implied by any MSM are (9) and that the restrictions on the distribution of $\overline{Y}_{\overline{a}}$ given $V^{\dagger}$ specified by the MSM hold for the conditional distribution of the observable $\overline{Y}\left(C\right)$ conditional on $\left(\overline{A}\left(C\right), V^{\dagger}\right)$.

For reasons that will become clear below, we indicate with a "*" any expectations, probabilities or hazard functions computed under the assumption that (8) and (9) hold. For convenience, denote $\overline{A}\left(C\right)$ as $\overline{A}$. Thus, for our MSM models (1a) – (2c), (8) implies the association models

Model 1a: $E^*\left[Y \mid V^{\dagger}, \overline{A}\right] = g\left(\overline{A}, V^{\dagger}, \beta_0\right)$

Model 1b: $\eta\left\{E^*\left[Y \mid V^{\dagger}, \overline{A}\right]\right\} = g\left(\overline{A}, V^{\dagger}, \beta_0\right) + g^{\dagger}\left(V^{\dagger}\right)$.

Model 1c: $R\left(\beta_0\right) \coprod^* \overline{A} \mid V^{\dagger}$ where $R\left(\beta_0\right) \equiv R\left(\overline{A}, \beta_0\right)$.

Model 1d: $E^*\left[Y\left(m\right) \mid V^{\dagger}, \overline{A}\right] = g_m\left[\overline{A}\left(m-1\right), V^{\dagger}; \beta_0\right]$, $m = 1, \ldots, K+1$.

Model 2a: $\lambda_T^*\left[t \mid V^{\dagger}, \overline{A}\right] = \lambda_T^*\left[t \mid V^{\dagger}, \overline{A}\left(t^{-}\right)\right] = \lambda_0\left(t\right) \exp\left[r\left(\overline{A}\left(t^{-}\right), t, V^{\dagger}, \beta_0\right)\right]$.

Model 2b: $\lambda_T^*\left[t \mid V^{\dagger}, \overline{A}\right] = \lambda_T^*\left[t \mid V^{\dagger}, \overline{A}\left(t^{-}\right)\right] = \lambda_0\left(t \mid V^{\dagger}\right) \exp\left[r\left(\overline{A}\left(t^{-}\right), t, V^{\dagger}, \beta_0\right)\right]$.

Model 2c: $\lambda_{R\left(\beta_0\right)}^*\left[u \mid V^{\dagger}, \overline{A}\right] = \lambda_{R\left(\beta_0\right)}^*\left[u \mid \overline{A}\left[r^{-1}\left(u, \overline{A}, V^{\dagger}, \beta_0\right)\right], V^{\dagger}\right] = \lambda_0\left(u \mid V^{\dagger}\right)$ where $R\left(\beta_0\right) \equiv R\left(\overline{A}, \beta_0\right)$ and $r^{-1}\left(u, \overline{a}, V^{\dagger}, \beta\right) \equiv t$ if $r\left(t, \overline{a}, V^{\dagger}, \beta\right) = u$.

We shall now consider estimation of these models for the observables, under assumption (9), and the further assumption that

$$\overline{A}\left(C\right) \textit{ has a known conditional distribution given } V^{\dagger} \ . \tag{10}$$

Semiparametric inference in the association models 1a - 2c without (10) imposed has been examined previously by many authors. Below we use their results to solve the estimation problem in our semiparametric model.

We will show that associated with each MSM model with (9) and (10) imposed is a class of regular asymptotically linear (RAL) estimators $\left\{\widehat{\beta}^*\left(h, \phi\right)\right\}$ for β_0, indexed by vector functions $h \in \mathcal{H}$ and

$\phi \in \Phi$ such that the set $\mathcal{IF}^* = \{IF^*(h,\phi)\}$ of influence functions of the $\widehat{\beta}^*(h,\phi)$ constitute all the influence functions for the model, in the sense that if $\widetilde{\beta}^*$ is any other RAL estimator, then the influence function of $\widetilde{\beta}^*$ equals $IF^*(h,\phi)$ for some functions $h \in \mathcal{H}, \phi \in \Phi$. Recall that an estimator $\widetilde{\beta}$ of β_0 is RAL with influence function IF if $n^{\frac{1}{2}}\left(\widetilde{\beta}-\beta_0\right) = n^{-\frac{1}{2}}\sum_i IF_i + o_p(1)$, the IF_i are i.i.d, and the convergence of $\widetilde{\beta}$ to β_0 is locally uniform. Here $o_p(1)$ denotes a random variable converging in probability to zero. Thus a RAL estimator is asymptotically equivalent to a sum of the i.i.d random variables IF_i. We obtain $\widehat{\beta}^*(h,\phi)$ by solving the estimating equations $n^{-\frac{1}{2}}\sum_i \widehat{D}_i^*(\beta,h,\phi) = o_p(1)$ described below. [We put $o_p(1)$ on the right side of the estimating equation to take care of cases (e.g., rank estimators) in which the estimating function $\widehat{D}^*(\beta,h,\phi)$ is not continuous in β and, thus, the left-hand side of the previous equality may never be exactly zero.] The solution $\widehat{\beta}^*(h,\phi)$ has influence function $IF^*(h,\phi) = \{\kappa^*(h)\}^{-1}U^*(\beta_0,h,\phi)$ where $U_i^*(\beta_0,h,\phi)$ depends only on subject i's data, $\kappa^*(h) = -\partial E^*[U^*(\beta,h,\phi)]/\partial\beta_{|\beta=\beta_0}$ does not depend on ϕ, and $n^{-\frac{1}{2}}\sum_i \widehat{D}_i^*(\beta_0,h,\phi) = n^{-\frac{1}{2}}\sum_i U_i^*(\beta_0,h,\phi) + o_p(1)$. Furthermore, $\Lambda^{\perp} = \{U^*(\beta_0,h,\phi)\}$ with $h \in \mathcal{H}, \phi \in \Phi$ is the linear span of $\mathcal{IF}^*$ and thus is the orthogonal complement to the nuisance tangent space for the model in the Hilbert space induced by the covariance norm. (Here we are quoting a well known result from the theory of semiparametric models. See Robins and Ritov (1997) for discussion.) We refer to $U^*(\beta_0,h,\phi)$ as the influence function for the estimating function $\widehat{D}(\beta_0,h,\phi)$. More specifically, $U^*(\beta,h,\phi)$ and $\widehat{D}^*(\beta,h,\phi)$ are each expressed as the sum of the two components, one of which $U_{tp}^*(\phi) = D_{tp}^*(\phi)$ is independent of the choice of the MSM and follows from the fact that, for the "treatment process (tp)," (9) and (10) are assumed. Specifically, if the $A(t)$ can jump only at times $0,1,2,\ldots$, $U_{tp}^*(\phi) =_{k=0}^{int(C)} \phi\left(k,\overline{A}(k),\overline{L}(k^-)\right) - E^*\left[\phi\left(k,\overline{A}(k),\overline{L}(k^-)\right) \mid \overline{L}(k^-),\overline{A}(k^-)\right]$ where $int(C)$ is the greatest integer less that or equal to C. It is easy to see that $\left\{U_{tp}^*(\phi)\right\}$ is, as ϕ varies, the sum over k of functions of the observed data $\left(\overline{A}(k),\overline{L}(k^-)\right)$ with mean zero given $\left(\overline{A}(k^-),\overline{L}(k^-)\right)$. If $A(t)$ is a continuous time marked point process, then $U_{tp}^*(\phi) = \int dM_A^*(u)\,\phi_1\left(u,\overline{A}(u^-),\overline{L}(u^-)\right) + \int dN_A(u)\left\{\phi_2\left(u,\overline{A}(u),\overline{L}(u^-)\right) - E^*[\phi_2\left(u,\overline{A}(u),\overline{L}(u^-)\right) \mid A(u) \neq A(u^-),\overline{L}(u^-),\overline{A}(u^-)]\right\}$ where $dM_A^*(u) = dN_A(u) - \lambda_A^*\left[u \mid \overline{A}(u^-),\overline{L}(u^-)\right]du$ and $dN_A(u) = I\{A(u) \neq A(u^-)\}$ counts jumps in the $\overline{A}$ process. [In the examples of the Introduction, we chose the function ϕ to be identically zero so that $\widehat{D}_{tp}^*(\phi)$ was also zero. As we shall see later, the choice ϕ identically zero, although computationally convenient because we can then use standard software, is somewhat inefficient.]

The other structural model-specific component $\widehat{D}_{sm}^*(\beta,h)$ and $U_{sm}^*(\beta,h)$ of $\widehat{D}^*(\beta,h,\phi)$ and $U^*(\beta,h,\phi)$ are the well-known estimating functions and their associated influence functions for the association models 1a - 2c with neither (9) nor (10) not imposed.

Model 1a: $\widehat{D}_{sm}^*(\beta,h) = U_{sm}^*(\beta,h) = h\left(\overline{A},V^\dagger\right)\varepsilon(\beta)$ with $\varepsilon(\beta) = Y - g\left(\overline{A},V^\dagger,\beta\right)$ and $h\left(\overline{A},V^\dagger\right)$ is any $\dim(\beta)$ vector function. In the linear logistic cumulative treatment model of the Introduction, $\widehat{D}_{sm}^*(\beta,h)$ was the score equation from the logistic model and thus $h\left(\overline{A},V^\dagger\right)$ was the vector $(1,cum(\overline{a}))'$.

Model 1b: $\eta(x) = x : \widehat{D}_{sm}^*(\beta,h) = U_{sm}^*(\beta,h) = \left\{\varepsilon(\beta) - h_1\left(\overline{A},V^\dagger\right)\right\}\left\{h_2\left(\overline{A},V^\dagger\right) - E^*\left[h_2\left(\overline{A},V^\dagger\right) \mid V^\dagger\right]\right\}$ where h_1 is any real valued function, $\varepsilon(\beta)$ is as just defined and the range of h_2 is of $\dim(\beta)$.

$\eta(x) = \ln[x/(1-x)]$: $U_{sm}^*(\beta,h) = U^\dagger(h,P(\beta))$ and $\widehat{D}_{sm}^*(\beta,h) \equiv U^\dagger\left(h,\widehat{P}(\beta)\right)$, where $P(\beta) = expit\left[g\left(\overline{A},V^\dagger,\beta\right) + g^\dagger\left(V^\dagger\right)\right]$, $\widehat{P}(\beta) = expit\left[g\left(\overline{A},V^\dagger,\beta\right) + \widehat{g}^\dagger\left(V^\dagger\right)\right]$, $expit(x) = e^x/(1+e^x)$, $\widehat{g}^\dagger\left(V^\dagger\right)$ is a $n^{\frac{1}{4}}$ - consistent estimate of $g^\dagger\left(V^\dagger\right)$, and $U^\dagger(h,P(\beta)) \equiv$
$\{Y - P(\beta)\}\left\{\left\{h\left(\overline{A},V^\dagger\right) - E^*\left[h\left(\overline{A},V^\dagger\right)P(\beta)\{1-P(\beta)\} \mid V^\dagger\right]/E^*\left[P(\beta)\{1-P(\beta)\} \mid V^\dagger\right]\right\}\right.$.

Model 1c: $\widehat{D}_{sm}^*(\beta,h) = U^*(\beta,h) = h\left[R(\beta),\overline{A},V^\dagger\right] - \int h\left[R(\beta),\overline{a},V^\dagger\right]dF^*\left[\overline{a} \mid V^\dagger\right]$.

Model 1d: Let $\varepsilon(\beta) = \{\varepsilon_1(\beta),\ldots\varepsilon_{K+1}(\beta)\}'$, $\varepsilon_m(\beta) = Y(m) - g_m\left[\overline{A}(m-1),V^\dagger;\beta\right]$. Then $\widehat{D}_{sm}^*(\beta,h) = U_{sm}^*(\beta,h) = h\left(\overline{A},V^\dagger\right)\varepsilon(\beta)$ where $h\left(\overline{A},V^\dagger\right)$ is now any $\dim(\beta) \times (K+1)$ matrix of real valued functions.

Model 2a: $\widehat{D}^*_{sm}(\beta,h) =_0^\infty dN_T(u)\left\{h\left(u,\overline{A}(u),V^\dagger\right) - \widetilde{\mathcal{L}}(h,u,\beta)\right\}$, where $\widetilde{\mathcal{L}}(h,u,\beta) = \widetilde{J}[h,\beta]/\widetilde{J}[\mathbf{1},\beta]$; for any $h\left(u,\overline{A}(u),V^\dagger\right), \widetilde{J}(h,\beta) = \widetilde{E}[h\left(u,\overline{A}(u),V^\dagger\right)I(T>u)\exp\left\{r\left[\overline{A}(u),u,V^\dagger,\beta\right]\right\}]$; for any H_i, $\widetilde{E}(H) =_{i=1}^n H_i/n$; $\mathbf{1}$ is the constant function equal to one; and $N_T(u) = I(T \le u)$. $U^*_{sm}(\beta,h) =_0^\infty dM_T(u)\left\{h\left(u,\overline{A}(u),V^\dagger\right) - \mathcal{L}^*\left(h,u,V^\dagger,\beta\right)\right\}$ where $\mathcal{L}^*(h,u,\beta) = J^*[h,\beta]/J^*[\mathbf{1},\beta]$; $J^*[h,\beta]$ is defined like $\widetilde{J}(h,\beta)$ but with E^* replacing $\widetilde{E}$; and $dM_T(u) = dN_T(u) - \lambda_T\left(u \mid \overline{A},V^\dagger\right)I(T>u)\,du$.
Model 2b: $U^*_{sm}(\beta,h)$ and $\widehat{D}^*_{sm}(\beta,h)$ are as above except $J^*(h,\beta) \equiv$
$E^*[h\left(u,\overline{A}(u),V^\dagger\right)I(T>u)\exp\left\{r\left[\overline{A}(u),u,V^\dagger,\beta\right]\right\} \mid V^\dagger]$ *and* $\widetilde{J}(h,\beta)$ replaces $E^*\left(\cdot \mid V^\dagger\right)$ in $J^*(h,\beta)$ by a $n^{\frac{1}{4}}$ - consistent estimator $\widehat{E}\left(\cdot \mid V^\dagger\right)$.
Model 2c: $\widehat{D}^*_{sm}(\beta,h) = \int_0^\infty du I\left[R(\beta)>u\right]\left\{H_2(u,\beta) - E^*\left[H_2(u,\beta) \mid V^\dagger\right]\right\} +$
$\int_0^\infty dN_{R(\beta)}(u)\left[H_1(u,\beta) - E^*\left[H_1(u,\beta) \mid V^\dagger\right]\right]$ and, for $j=0,1$, $H_j(u,\beta) = h_j\left[u,\overline{A}\left\{r^{-1}\left(u,\overline{A},V^\dagger,\beta\right)\right\},V^\dagger\right]$.
$U^*_{sm}(\beta,h) = \widehat{D}^*_{sm}(\beta,h) - E^*\left[D^*_{sm}(\beta,h) \mid \overline{A},V^\dagger\right]$.

Remark: Note that in model 2b and in model 1b with $\eta(x) = \ln\left[x/(1-x)\right]$, smooths are necessary to estimate $g^\dagger\left(V^\dagger\right)$ and $E^*\left(\cdot \mid V^\dagger\right)$ if $V^\dagger$ has continuous components. In particular, due to the curse of dimensionality, it is not possible to obtain a reasonable $n^{\frac{1}{2}}$ – consistent estimator of β_0 in these models when $V^\dagger$ has multiple continuous components. This can be formalized using the concept of curse of dimensionality appropriate (CODA) semiparametric information bounds introduced by Robins and Ritov (1997). Specifically, models 2b and 1b have CODA information bounds of zero, although they have positive ordinary semiparametric information bounds.

3.2 Non-ancillary treatment process

In this section, we no longer assume (8) is true. The essential idea of this section (requiring some minor modification) is to reweight $\widehat{D}^*_{sm}(\beta,h)$ by the inverse of a subject's probability of having had his observed treatment history. We continue to assume that

$$f\left[a(t) \mid \overline{L}\left(t^-\right),\overline{A}\left(t^-\right),V^\dagger\right] \quad \text{is known for } t \le C \tag{11}$$

which implies that if $A(t)$ jumps at non-random times $0,\ldots,K$, $W(k) = f\left[A(k) \mid \overline{L}\left(k^-\right),\overline{A}\left(k^-\right)\right]$ and $\overline{W}(k) =_{m=0}^k W(k)$ are known. If $A(t)$ jumps in continuous time,$\overline{W}(t) =$
$\exp\left[-\int_0^t \lambda_A\left[u \mid \overline{L}\left(u^-\right),\overline{A}\left(u^-\right)\right]du\right]_{\{u;A(u)\neq A(u^-),u<t\}} \lambda_A\left[u \mid \overline{L}\left(u^-\right),\overline{A}\left(u^-\right)\right]$
$f[A(u) \mid \overline{A}\left(u^-\right),\overline{L}\left(u^-\right),A(u) \neq A\left(u^-\right)]$ is known.

We now come to a subtle but crucial idea. We need to artificially censor a subject at the first time $C^\dagger$ that the density of receiving his observed treatment $A\left(C^\dagger\right)$ at $C^\dagger$ was zero for some prognostic factor history $\overline{\ell}\left(C^{\dagger -}\right)$ in order to insure that the reweighted $\widehat{D}^*_{sm}(\beta_0,h)$ still has asymptotic mean zero. We formalize this idea as follows. If $A(t)$ jumps at non-random times, let $\mathring{\mathcal{A}}\left(k,\overline{a}\left(k^-\right),v^\dagger\right) = \left\{a(k); f\left[a(k) \mid \overline{L}\left(k^-\right),\overline{A}\left(k^-\right) = \overline{a}\left(k^-\right),V^\dagger = v^\dagger\right] \neq 0 \ w.p.1\right\}$ and set $C^\dagger =$
$\min\left\{k; A(k) \notin \mathring{\mathcal{A}}\left(k,\overline{A}\left(k^-\right),V^\dagger\right)\right\}$. If $A(t)$ jumps in continuous time, let $\mathring{\mathcal{A}}\left(t,\overline{a}\left(t^-\right),v^\dagger\right) = \{a(t); f[a(t) \mid \overline{L}\left(t^-\right),\overline{A}\left(t^-\right) = \overline{a}\left(t^-\right),A(t) \neq a\left(t^-\right),V^\dagger = v^\dagger] \neq 0 \ w.p.1$ or $a(t) = A\left(t^-\right)\}$ and set $C^\dagger =$
$\inf\left\{t; A(t) \notin \mathring{\mathcal{A}}\left(t,\overline{A}\left(t^-\right),V^\dagger\right)\right\}$. The variable $C^\dagger$ is crucial because, as indicated in the remark following Lemma 3.1 below, one can only unbiasedly reweight a function of $A(t)$ for $A(t) \in \mathring{\mathcal{A}}\left(t,\overline{a}\left(t^-\right),v^\dagger\right)$.

Let $f^*\left(\overline{a} \mid V^\dagger\right)$ be a density (chosen by the analyst). Let F^* denote the joint distribution which differs from the true distribution F of O only in that $f\left[a(t) \mid \overline{L}\left(t^-\right),\overline{A}\left(t^-\right),V^\dagger\right]$ is replaced by the ancillary density $f^*\left[a(t) \mid \overline{A}\left(t^-\right),V^\dagger\right]$. Further, define $\mathcal{W}(t) \equiv \overline{W}(t)/f^*\left[\overline{A}(t) \mid V^\dagger\right]$ and $O^\dagger = \left(\overline{L}\left(C^\dagger\right),\overline{A}\left(C^\dagger\right)\right)$.

Note that in the examples of the Introduction, we chose $f^*\left[a(t) \mid \overline{A}(t^-), V^\dagger\right]$ to be $f\left[a(t) \mid \overline{A}(t^-), V^\dagger\right]$. Note even with this choice, the distribution F of O differs from the distribution F^* if (9) is false. A key result, which follows from direct calculation, is

Lemma 3.1: For any $z(O^\dagger)$, $E\left[z(O^\dagger)/\mathcal{W}(C^\dagger) \mid V^\dagger\right] = E^*\left[z(O^\dagger) \mid V^\dagger\right]$.

Remark: It is false that $E\left[z(O)/\mathcal{W}(C) \mid V^\dagger\right] = E^*\left[z(O) \mid V^\dagger\right]$.

Let $D^*_{sm}(\beta,h)$ be the probability limit under F^* of $\widehat{D}^*_{sm}(\beta,h)$ and let $\{U_{sm}(\beta,h)\}$ and $\{D_{sm}(\beta,h)\}$ be the subsets of $\{U^*_{sm}(\beta,h)\}$ and $\{D^*_{sm}(\beta,h)\}$, respectively, that depend on the data only through $O^\dagger$. Set $\mathcal{W} = \mathcal{W}(C^\dagger)$ and note $D_{sm}(\beta,h)$ and $U_{sm}(\beta,h)$ often depend on $E^*\left[\cdot \mid V^\dagger\right] = E\left[\cdot/\mathcal{W} \mid V^\dagger\right]$ or $E^*[\cdot] = E[\cdot/\mathcal{W}]$. Define $\widehat{D}_{sm}(\beta,h)$ and $\widehat{U}_{sm}(\beta,h)$ like $D_{sm}(\beta,h)$ and $U_{sm}(\beta,h)$ except replace any unknown expectations $E\left[\cdot/\mathcal{W} \mid V^\dagger\right]$ and $E[\cdot/\mathcal{W}]$ with appropriate estimates $\widehat{E}\left[\cdot/\mathcal{W} \mid V^\dagger\right]$ and $\widehat{E}[\cdot/\mathcal{W}]\cdot$.

Examples: In model 1b, with $\eta(x) = x$, $U_{sm}(\beta,h) = \widehat{D}_{sm}(\beta,h) = U^*_{sm}(\beta,h)$ has $h_1\left(\overline{A}, V^\dagger\right)$ and $h_2\left(\overline{A}, V^\dagger\right)$ being functions only of $\left\{\overline{A}(C^\dagger), V^\dagger\right\}$. Note $E^*\left[h_2\left(\overline{A}, V^\dagger\right) \mid V^\dagger\right]$ is known and need not be estimated.

In contrast, in models 2a and 2b, $\widehat{D}_{sm}(\beta,h)$ will be defined like $\widehat{D}^*_{sm}(\beta,h)$ except in defining $\widetilde{J}(h,\beta)$ we replace $I(T>u)$ by $I(T>u)/\mathcal{W}$ in order to estimate the unknown expectations. Alternatively, we can replace $I(T>u)$ by $I(T>u)/\mathcal{W}(u)$. This latter choice *(i)* will in general have better finite sample properties, *(ii)* tend to increase efficiency unless the estimator with $I(T>u)/\mathcal{W}$ was already semiparametric efficient, and *(iii)* was the approach we took in the Introduction. The issues are exactly those discussed in Robins (1993), which the reader may consult for further clarification.

In model 1b with $\eta(x) = \ln[x/(1-x)]$, $\widehat{g}\left(V^\dagger\right) \equiv \widehat{g}\left(V^\dagger, \beta\right)$ could be chosen to minimize $\widetilde{E}\left[\left(Y - expit\left\{g\left(\overline{A}, V^\dagger, \beta\right) + g^\dagger\left(V^\dagger\right)\right\}\right)^2/\mathcal{W}\right]$ over $g^\dagger\left(V^\dagger\right)$ in some class (e.g., splines), whose dimension may increase with sample size.

In the appendix we briefly sketch a proof of the following.

Theorem 3.1: Subject to regularity conditions, in the semiparametric model *(i)* characterized by (4), (11), the data O, and a MSM, the class $\left\{\widehat{\beta}(h,\phi)\right\}$ with $h \in \mathcal{H}$ and $\phi \in \Phi$ of estimators which solve $0 = \sum_i \widehat{D}_i(\beta,h,\phi)$ with $\left\{\widehat{D}(\beta,h,\phi)\right\} = \left\{\widehat{D}_{sm}(\beta,h)/\mathcal{W} + D_{tp}(\phi)\right\}$ is a class of RAL estimators with influence functions $\mathcal{IF} = \{IF(h,\phi)\}$, $IF(h,\phi) = \{\kappa(h)\}^{-1} U(\beta_0,h,\phi)$, $\kappa(h) = -\partial E[U(\beta,h,\phi)]/\partial\beta_{|\beta=\beta_0}$, $U(\beta_0,h,\phi) = U_{sm}(\beta_0,h)/\mathcal{W} + U_{tp}(\phi)$, where $U_{tp}(\phi) = D_{tp}(\phi)$ is defined like $U^*_{tp}(\phi)$ except with the true law F replacing F^*. Furthermore, $\mathcal{IF}$ is the set of all influence functions.

3.3 Efficiency for fixed h

We now begin to explore efficiency issues. By a projection argument similar to that given in Robins et al. (1994), we have

Theorem 3.2: For a given h, among all estimators $\widehat{\beta}(h,\phi)$, the most efficient has ϕ equal to $\phi_{opt} \equiv \phi_{opt}(h)$: if $A(t)$ only jumps at non-random times $0, 1, 2, \ldots$ then $\phi_{opt} \equiv 0$ if $k > C^\dagger$, and if $k \le C^\dagger$, $\phi_{opt}\left[k, \overline{a}(k), \overline{\ell}(k^-)\right] = E[U_{sm}(h)/\mathcal{W} \mid \overline{A}(k) = \overline{a}(k), \overline{L}(k^-) = \overline{\ell}(k^-)] = \left\{\overline{W}(k)\right\}^{-1} \iint d\mu\left(\underline{a}_{k+1}\right) f^*\left(\overline{a} \mid v^\dagger\right)$ $E[u_{sm}\left\{\overline{a}(C^\dagger), \overline{Y}_{\overline{a}}(C^\dagger), V^\dagger, h\right\} \mid \overline{L}_{\overline{a}}(k^-) = \overline{\ell}(k^-), \overline{A}(k) = \overline{a}(k)]$ and $U_{sm}(h) \equiv U_{sm}(\beta_0,h)$. Furthermore, if CAR holds, $\overline{A}(k) = \overline{a}(k)$ can be removed from the last conditioning event above. If $A(t)$ jumps in continuous time, $\phi_{1,opt} = E[U_{sm}(h)/\mathcal{W} \mid \overline{L}(u^-), \overline{A}(u^-), A(u) \neq A(u^-)] - E\left[U_{sm}(h)/\mathcal{W} \mid \overline{L}(u^-), \overline{A}(u^-)\right]$ since $E\left[U_{sm}(h)/\mathcal{W} \mid \overline{L}(u^-), \overline{A}(u^-)\right] = E\left[U_{sm}(h)/\mathcal{W} \mid \overline{L}(u^-), \overline{A}(u^-), A(u) = \overline{A}(u^-)\right]$, and $\phi_{2,opt} = E\left[U_{sm}(h)/\mathcal{W} \mid \overline{A}(u), \overline{L}(u^-)\right]$.

We now relax unrealistic assumption (11) that the conditional density of $A(t)$ is known. We consider two cases. In the first case the density is completely unknown and in the second the density follows a parametric model.

Theorem 3.3: a) The semiparametric model *(ii)* characterized by (4), data O, and a MSM (with (11) not imposed), has the set of influence functions $\{IF(h, \phi_{opt}(h))\}$ with $h \in \mathcal{H}$.
b): In model *(iii)* characterized by (4), data O, a MSM, and a parametric model indexed by parameter α for $f\left[a(t) \mid \overline{L}(t^-), \overline{A}(t^-)\right]$, the set of influence functions for the model is the set $\left\{\kappa(h)^{-1}\left[U(h,\phi) - E\left[U(h,\phi) S'_\alpha\right]\left\{E\left[S_\alpha S'_\alpha\right]\right\}^{-1} S_\alpha\right]\right\}$ of influence functions of $\left\{\widehat{\beta}(h,\phi,\widehat{\alpha})\right\}$ solving $o_p(1) = n^{-\frac{1}{2}} \sum_i \widehat{D}_i(\beta, h, \phi, \widehat{\alpha})$ where $\widehat{\alpha}$ is the MLE of α, $S_\alpha = \partial \log\left\{f\left[A(t) \mid \overline{L}(t^-), \overline{A}(t^-), \alpha\right]/\partial\alpha\right\}$ is the subject-specific score for α, and $D(\beta, h, \phi, \widehat{\alpha})$ is $D(\beta, h, \phi)$ evaluated at $\widehat{\alpha}$.

Theorem 3.3b can be extended to semiparametric models for $f\left(a(t) \mid \overline{L}(t^-), \overline{A}(t^-)\right)$ such as a Cox proportional hazard model as in Robins (1993; 1998b, App. 2). In the non- and semi-parametric case, we have to plug an estimator of $f\left[a(t) \mid \overline{L}(t^-), \overline{A}(t^-)\right]$ into the estimating function $\widehat{D}_i(\beta, h, \phi)$. In general, the estimator needs to converge to the true density at a rate greater than $n^{\frac{1}{4}}$ to obtain a RAL estimator of β.

3.4 Censoring

No new idea is required to account for and adjust for right censoring. Specifically, let Q be censoring time in MSMs. Define a censoring process $A_2(u)$ by $A_2(u) = 0$ if $Q > u$ and $A_2(u) = 1$ otherwise. Let the treatment of interest be $a_1(u)$ and define $a(u) = (a_1(u), a_2(u))$ and write $T_{\overline{a}}$ as $T_{\overline{a}_1, \overline{a}_2}$. To want to adjust for censoring is only to say that interest is in the direct effect of $\overline{a}_1$ when $\overline{a}_2 \equiv 0$, i.e., when censoring is abolished. As a concrete example, the Cox model MSM 2a in the presence of censoring would become

$$\lambda_{T_{\overline{a}_1, \overline{a}_2 \equiv 0}}\left(t \mid V^\dagger\right) = \lambda_0(t) \exp\left\{r\left[\overline{a}_1(t^-), t, V^\dagger; \beta_0\right]\right\}.$$

If $\overline{A}$ is ancillary (now including the censoring process), $\widehat{D}^*_{sm}(\beta, h)$ and $U^*_{sm}(\beta, h)$ are as above except that now $N_T(u) = I[T \le u]\, I[T < Q]$ and $I[T > u]$ is everywhere replaced by $I[T > u]\, I[Q > u]$. Of course $\overline{W}(t)$ is now the probability that a subject would have his observed treatment and censoring history. This is exactly the approach we took in analyzing the 002 trial data in the Introduction.

4 Semiparametric Efficiency

4.1 The efficient score

In any semiparametric model, the semiparametric variance bound is the inverse of the variance of the efficient score S_{eff}. The efficient score in models *(i)* - *(iii)* of Theorems 3.1 and 3.2 are the same and, by Theorem 5.3 in Newey and McFadden (1993), equal $S_{eff} = U\left(\beta_0, h_{eff}, \phi_{eff}\right)$ where $\phi_{eff} = \phi_{opt}(h_{eff})$ and h_{eff} is uniquely characterized by the requirement that for all $U(\beta_0, h, \phi)$

$$E\left[U(\beta_0, h, \phi)\, U\left(\beta_0, h_{eff}, \phi_{opt}(h_{eff})\right)'\right] = \kappa(h)$$

which is equal to

$$E\left[U_{sm}(\beta_0, h)\, U\left(\beta_0, h_{eff}, \phi_{opt}(h_{eff})\right)'\right] = \kappa(h) \quad . \tag{12}$$

To show how to use (12) to calculate h_{eff}, we consider the following simple example.

Model 1a: Consider MSM 1a with $C^\dagger = C = K + 1 = 1$ w.p.1 so $K = 0$ and the $\overline{A}(C)$ process only jumps at time zero. So $\overline{A} = A(0)$ and $\mathcal{W}^{-1} = f^*\left[\overline{A} \mid V^\dagger\right] / f\left[\overline{A} \mid L(0)\right]$ where $V^\dagger \subset L(0) = V(0)$. For the purposes of computing the efficient score, we can choose $f^*\left(\overline{A} \mid V^\dagger\right)=1$ w.p.1 without worrying that it is not a density, because it can be absorbed into $h_{eff}\left(\overline{A}, V^\dagger\right)$. In Appendix 2, we prove the following.

Theorem 4.1: With $f^*\left(\overline{A} \mid V^\dagger\right) = 1$ w.p.1, Eq. 12 implies $h_{eff}\left(\overline{A}, V^\dagger\right)$ is the unique solution to the type two Fredholm equation $h_{eff}\left(\overline{A}, V^\dagger\right)\left[\int var\left[\varepsilon \mid \overline{A}, L(0)\right]\left\{f\left(\overline{A} \mid L(0)\right)\right\}^{-1} f\left(V^\bullet \mid V^\dagger\right) d\mu\left(V^\bullet\right)\right]+$ $\int h_{eff}\left(\overline{a}, V^\dagger\right)\omega\left(\overline{a}, \overline{A}, V^\dagger\right) d\mu\left(\overline{a}\right)=\partial g\left(\overline{A}, V^\dagger, \beta_0\right)/\partial\beta$ where $V^\bullet = L(0)\backslash V^\dagger$ and $\omega\left(\overline{a}, \overline{A}, V^\dagger\right) =$ $\left[\int E\left[\varepsilon \mid \overline{a}, L(0)\right] E\left[\varepsilon \mid \overline{A}, L(0)\right] f\left(V^\bullet \mid V^\dagger\right) d\mu\left(V^\bullet\right)\right]$. Note that if $\overline{A}$ has finite support, this is a finite dimensional matrix equation. Our estimators, specialized to this example, are continuous treatment extensions of efficient propensity score estimators of an average treatment effect. By dividing by the propensity score, we eliminate the bias due to within stratum confounding that can occur with subclassification on the propensity score as recommended by Rosenbaum and Rubin (1983)

4.2 Efficiency calculations using missing data theory

Given (4), imposing CAR cannot change the efficient score. Thus, it is of interest to rederive the efficient score using the Hilbert space results of van der Vaart (1991) and of Robins et al. (1994) for missing data models under CAR. For convenience, assume C is non-random and write $\overline{A} \equiv \overline{A}(C)$. The full data are $\overline{L}^F = \left\{\overline{L}_{\overline{a}}; \overline{a} \in \mathcal{A}\right\}$. Given any $B = b\left(\overline{L}^F\right)$, the score operator $\mathbf{s}(B) = E[B \mid O], O = \left(\overline{A}, \overline{L}_{\overline{A}}\right)$. For any $Q = q(O)$, the non-parametric adjoint operator $\mathbf{s}^\dagger$ under CAR is $\mathbf{s}^\dagger(Q) = E\left[Q \mid \overline{L}^F\right] =$ $\int d\mu\left(\overline{a}\right) q\left(\overline{a}, \overline{L}_{\overline{a}}\right) f\left[\overline{a} \mid \overline{L}_{\overline{a}}\right]$. Suppose for the remainder of this subsection that the $\overline{A}$ jumps only at times $0, 1, \ldots, K$ and $\overline{L}$ jumps at $0^-, \ldots, K+1^-$ and $C = K+1$. We then have by CAR

$$f\left[\overline{a}(k) \mid \overline{L}_{\overline{a}}\right] = \prod_{m=0}^{k} f\left[a(m) \mid \overline{a}(m-1), \overline{L}_{\overline{a}}(m)\right] \tag{13}$$

and $f\left[\overline{a} \mid \overline{L}_{\overline{a}}\right] = f\left[\overline{a}(K) \mid \overline{L}_{\overline{a}}\right]$. It is then easy to check the null space of $\mathbf{s}^\dagger$, $N\left(\mathbf{s}^\dagger\right) = \left\{U_{tp}(\phi)\right\}$. Now define the non-parametric information operator, $\mathbf{m} = \mathbf{s}^\dagger\mathbf{s} : \overline{L}^F \to R\left(\mathbf{s}^\dagger\right)$ where $R\left(\mathbf{s}^\dagger\right)$ is the range of $\mathbf{s}^\dagger$. Note that $R\left(\mathbf{s}^\dagger\right) = \left\{B = \int d\mu\left(\overline{a}\right) b\left(\overline{a}, \overline{L}_{\overline{a}}\right)\right\}$. Let $\mathbf{m}^{-1} : R\left(\mathbf{s}^\dagger\right) \to R\left(\mathbf{s}^\dagger\right)$ be the inverse of $\mathbf{m}$ on $R\left(\mathbf{s}^\dagger\right)$. Given $\overline{a}_1, \overline{a}_2$, let u_{12} be the smallest u with $a_1(u) \neq a_2(u)$. We then have by a direct calculation
Theorem 4.2: If for all $\overline{a}_1, \overline{a}_2$

$$\overline{L}_{\overline{a}_1} \coprod \overline{L}_{\overline{a}_2} \mid \overline{L}_{\overline{a}_1}\left(u_{12}^-\right) \tag{14}$$

then

$$\mathbf{m}^{-1}\left[\int d\mu\left(\overline{a}\right) b\left(\overline{a}, \overline{L}_{\overline{a}}\right)\right] = \int d\mu\left(\overline{a}\right) \left\{_{m=1}^{K+1} \left\{f\left[\overline{a}(m-1) \mid \overline{L}_{\overline{a}}\right]\right\}^{-1}\right. \tag{15}$$
$$\left\{E\left[b\left(\overline{a}, \overline{L}_{\overline{a}}\right) \mid \overline{L}_{\overline{a}}(m)\right] - E\left[b\left(\overline{a}, \overline{L}_{\overline{a}}\right) \mid \overline{L}_{\overline{a}}(m-1)\right]\right\} + E\left[b\left(\overline{a}, \overline{L}_{\overline{a}}\right) \mid \overline{L}_{\overline{a}}(0)\right]\right\} .$$

Remark: Gill and Robins (1999) show that (14) places no restriction on the law of the observed data O even when sequential randomization and a MSM are imposed. We can and do always assume that (14) holds.

Now let S_{eff}^F and $\Lambda^{F,\perp}$ be the efficient score and the orthogonal complement to the nuisance tangent space for the parameter β of our marginal structural model when we have data on $\overline{L}^F$. Then the efficient score S_{eff} based on data O under CAR is $g\left[m^{-1}\left(D_{eff}\right)\right]$ where D_{eff} is the unique member of $\Lambda^{F,\perp} \cap R\left(\mathbf{s}^\dagger\right)$ satisfying

$$\Pi\left[\mathbf{m}^{-1}\left(D_{eff}\right) \mid \Lambda^{F,\perp}\right] = S_{eff}^F \ , \tag{16}$$

where Π is the Hilbert space projection operator. To show how to use this result to calculate S_{eff}, we revisit the example given in the last subsection.

Example: Model 1a: Consider MSM 1a as in Sec. 4.1. Then, by an extension of Theorem 8.3 of Robins et al. (1994)

$$\Lambda^{F,\perp} = \left\{ \int d\mu\left(\overline{a}\right) h\left(\overline{a}\right) \varepsilon\left(\overline{a}\right) \right\} \tag{17}$$

where *(i)* $\varepsilon\left(\overline{a}\right) = \varepsilon\left(\overline{a}, \beta_0\right)$, *(ii)* $h\left(\overline{a}\right)$ is a vector valued function of the dimension of β_0. Note that $\Lambda^{F,\perp}$ is contained in $R\left(\mathbf{s}^{\dagger}\right)$ as will be the case for MSMs with positive information.

Remark:

$$\textit{If } \overline{\mathcal{A}} = \{\overline{a}_1, \ldots, \overline{a}_S\} \textit{ is finite,} \tag{18}$$

then $\varepsilon\left(\overline{a}\right)$ can be identified with the S vector that has components $\varepsilon_s\left(\overline{a}_s\right) = Y_{\overline{a}_s} - g\left(\overline{a}_s, V^{\dagger}, \beta_0\right)$. For arbitrary $\overline{\mathcal{A}}$, $\varepsilon\left(\overline{a}\right)$ is a stochastic process with index set $\overline{\mathcal{A}}$. For $\overline{a}, \overline{a}^* \in \overline{\mathcal{A}}$, let $\mathbf{cv}\left(\overline{a}, \overline{a}^*\right) = cov\left(\varepsilon\left(\overline{a}\right), \varepsilon\left(\overline{a}^*\right)\right)$. If $\overline{\mathcal{A}}$ is given by (18), $\mathbf{cv}\left(\overline{a}, \overline{a}^*\right)$ corresponds to the $S \times S$ matrix with j, k entry $\mathbf{cv}\left(\overline{a}_j, \overline{a}_k\right)$. Let $\mathbf{cv}^{-1}\left(\overline{a}^{**}, \overline{a}^*\right)$ be a (generalized) inverse of $\mathbf{cv}\left(\overline{a}, \overline{a}^*\right)$, i.e., by definition, for any function $q\left(\overline{a}^*\right)$, $\int \left[\int \mathbf{cv}^{-1}\left(\overline{a}^{**}, \overline{a}\right) \mathbf{cv}\left(\overline{a}, \overline{a}^*\right) d\mu\left(\overline{a}\right)\right] q\left(\overline{a}^*\right) d\mu\left(\overline{a}^*\right) = q\left(\overline{a}^{**}\right)$. In particular, if (18) holds, $\mathbf{cv}^{-1}\left(\overline{a}^*, \overline{a}\right)$ is just the inverse of the matrix identified with $\mathbf{cv}\left(\overline{a}, \overline{a}^*\right)$. Then, generalizing Chamberlain (1987),

$$S_{eff}^{F} = \int d\mu\left(\overline{a}\right) \left\{\partial g\left(\overline{a}, V^{\dagger}; \beta_0\right) / \partial \beta\right\} \left[\int \mathbf{cv}^{-1}\left(\overline{a}, \overline{a}^*\right) \varepsilon\left(\overline{a}^*\right) d\mu\left(\overline{a}^*\right)\right] . \tag{19}$$

If $\overline{\mathcal{A}}$ is given by (18), $\partial g\left(\overline{a}, V^{\dagger}; \beta_0\right) / \partial \beta$ can be identified with the $\dim \beta \times S$ matrix with j, k entry $\partial g\left(\overline{a}_k, V^{\dagger}; \beta_0\right) / \partial \beta_j$. Again, generalizing Theorem 8.3 in Robins et al. (1994),

$$\Pi\left[\int d\mu\left(\overline{a}\right) b\left(\overline{a}, \overline{L}_{\overline{a}}\right) \mid \Lambda^{F,\perp}\right] = \int d\mu\left(\overline{a}\right) E\left[b\left(\overline{a}, \overline{L}_{\overline{a}}\right) \varepsilon\left(\overline{a}\right) \mid V^{\dagger}\right] \tag{20}$$
$$\left[\int \mathbf{cv}^{-1}\left(\overline{a}, \overline{a}^*\right) \varepsilon\left(\overline{a}^*\right) d\mu\left(\overline{a}^*\right)\right] .$$

Hence to solve (16), we need to find the solution $h_{eff}\left(\overline{a}, V^{\dagger}\right)$ to the equation

$$E[\mathbf{m}^{-1}\left\{\int d\mu\left(\overline{a}^*\right) h\left(\overline{a}^*, V^{\dagger}\right) \varepsilon\left(\overline{a}^*\right)\right\} \varepsilon\left(\overline{a}\right) \mid V^{\dagger}] = \partial g\left(\overline{a}, V^{\dagger}; \beta_0\right) / \partial \beta . \tag{21}$$

By (15) with $K = 0$ and the fact that, by CAR, $f\left(\overline{a} \mid \overline{L}_{\overline{a}}\right) = f\left(\overline{a} \mid L(0)\right)$, the LHS of (21) can be written

$$E\{\int d\mu\left(\overline{a}^*\right) h\left(\overline{a}^*, V^{\dagger}\right) \{\left[\varepsilon\left(\overline{a}^*\right) - E\left[\varepsilon\left(\overline{a}^*\right) \mid L(0)\right]\right] \left\{f\left(\overline{a}^* \mid L(0)\right)\right\}^{-1} +$$
$$E\left[\varepsilon\left(\overline{a}^*\right) \mid L(0)\right] \}\varepsilon\left(\overline{a}\right) \mid V^{\dagger}\} = E\{\int d\mu\left(\overline{a}^*\right) h\left(\overline{a}^*, V^{\dagger}\right)$$
$$[cov\left[\varepsilon\left(\overline{a}^*\right), \varepsilon\left(\overline{a}\right) \mid L(0)\right] \left\{f\left(\overline{a}^* \mid L(0)\right)\right\}^{-1} + E\left[\varepsilon\left(\overline{a}^*\right) \mid L(0)\right] E\left[\varepsilon\left(\overline{a}\right) \mid L(0)\right]] | V^{\dagger}\} .$$

However, since by assumption (14), $Y_{\overline{a}_j} Y_{\overline{a}_k} \mid \bar{L}(0)$ for $k \neq j$, (21) reduces to

$$h\left(\overline{a}, V^{\dagger}\right) E\{var\left[\varepsilon\left(\overline{a}\right) \mid L(0)\right] f\left(\overline{a} \mid L(0)\right)^{-1} \mid V^{\dagger}\} +$$
$$\int d\mu\left(\overline{a}^*\right) h\left(\overline{a}^*, V^{\dagger}\right) E\{E\left[\varepsilon\left(\overline{a}^*\right) \mid L(0)\right] E\left[\varepsilon\left(\overline{a}\right) \mid L(0)\right] \mid V^{\dagger}\}$$
$$= \partial g\left(\overline{a}, V^{\dagger}, \beta_0\right) / \partial \beta .$$

Upon noting that, by CAR, $f\left[\varepsilon \mid \overline{A} = \overline{a}, L(0)\right] = f\left[\varepsilon\left(\overline{a}\right) \mid L(0)\right], w$. We see that this is the same expression for $h_{eff}\left(\overline{a}, V^{\dagger}\right)$ as obtained in Theorem 4.1.

4.3 A practical approach to obtaining reasonable efficiency

Estimation of h_{eff} is computationally difficult because of the need to solve integral equations without closed form solutions. A practical approach to choosing h and $f^*\left(\overline{a} \mid V^\dagger\right)$ is important. Given a model for $f\left[a(t) \mid \overline{a}(t^-), \overline{\ell}(t^-)\right]$ depending on parameter $\alpha' = (\alpha_1', \alpha_2')$ such that $\alpha_1 = 0 \Leftrightarrow f\left[a(t) \mid \overline{a}(t^-), \overline{\ell}(t^-)\right] = f\left[a(t) \mid \overline{a}(t^-), v^\dagger\right]$, rather than choosing $f^*\left[\overline{a} \mid V^\dagger\right]$, we use $f^*\left[\overline{a} \mid V^\dagger; \widetilde{\alpha}_2\right]$ where $\widetilde{\alpha}_2$ is the MLE of α_2 with α_1 set to zero. This is exactly the approach we took in analyzing the 002 data in the Introduction. [The fact that $f^*\left[\overline{a} \mid V^\dagger\right]$ is estimated does not influence the asymptotic distribution of $\widehat{\beta}(h, \phi)$.] It follows that if (8) holds [i.e., $\overline{A}$ is an ancillary process], $\mathcal{W}$ will converge to 1. Further, in each of the models 1a – 2c of Sec. 3.1, the efficient choice of h, say, h_{opt}, for solving $\sum_i \widehat{D}^*_{sm,i}(h) = 0$ when (9) is not imposed is well known. We suggest choosing h to be h_{opt} or an estimate $\widehat{h}_{opt}$ thereof, and choosing ϕ to be an estimate of $\phi_{opt}\left(\widehat{h}_{opt}\right)$. Such a choice guarantees that if $\overline{A}$ is an ancillary process, our estimate of β_0 will be more efficient than the estimate based on solving $0 = \sum_i \widehat{D}_{sm,i}(h_{opt})$. Specifically, in MSM 1a, $h_{opt} = \{\partial\varepsilon(\beta_0)/\partial\beta\}\left\{var\left(\varepsilon(\beta_0) \mid \overline{A}, V^\dagger\right)\right\}^{-1}$. For model 1b, Chamberlain (1988) gives $h_{1,opt}$ and $h_{2,opt}$. In model 1c, $h_{opt} = [\partial/\partial\beta]\left[ln\left\{[\partial R(\beta_0)/\partial Y] f\left[R(\beta_0) \mid V^\dagger\right]\right\}\right]$. In model 1d, h_{opt} is as in model 1a with $\varepsilon(\beta_0)$ now a vector. In model 2a, $h_{opt} = \partial lnr\left[\overline{A}(u^-), u, V^\dagger, \beta_0\right]/\partial\beta$. For model 2b, h_{opt} is given by Sasieni (1992). In model 2c, $h_{opt,2} = h_{opt,1}\lambda_0\left(u \mid V^\dagger\right)$ and $h_{opt,1} = \partial \ln\lambda_{R(\beta)}\left(u \mid V^\dagger\right)/\partial\beta_{|\beta=\beta_0}$.

5 Comparison of MSMs and SNMs

We begin by recalling the definition of a structural nested distribution model.

5.1 Structural Nested Distribution Models

For concreteness, we consider the setting of the MSM model 1a - 1c with $Y_{\overline{a}} = Y_{\overline{a}}(K+1)$ and the A process and L process jumping at non-random times $0, \ldots, K$ and $0^-, \ldots, K+1^-$ respectively. Henceforth, we take $V^\dagger = \emptyset$. Suppose Y is a continuous variable with a continuous distribution function $F_Y(y) = pr[Y < y]$. Let $(\overline{a}(m), 0)$ denote the treatment history given by $\overline{a}(m)$ through time m and zero at times $m+1, \ldots, K$. Then let $\gamma\left(y, \overline{\ell}(m), \overline{a}(m)\right)$ be the unique function mapping quantiles of $Y_{\overline{a}(m),0}$ into those of $Y_{\overline{a}(m-1),0}$ conditional on $\overline{L}(m) = \overline{\ell}(m), \overline{A}(m) = \overline{a}(m)$ so that $\gamma\left(y, \overline{\ell}(m), \overline{a}(m)\right)$ measures the magnitude of the effect of a final blip of treatment $a(m)$ on quantiles of Y among subjects with observed history $\left\{\overline{\ell}(m), \overline{a}(m)\right\}$. A structural nested distribution model (SNDM) is a parametric model for this function. That is, it specifies that $\gamma\left(y, \overline{\ell}(m), \overline{a}(m)\right) = \gamma\left(y, \overline{\ell}(m), \overline{a}(m), \beta_0\right)$ where $\gamma\left(y, \overline{\ell}(m), \overline{a}(m), \beta\right)$ is a known increasing function of y satisfying $\gamma\left(y, \overline{\ell}(m), \overline{a}(m), \beta\right) = y$ if $a(m) = 0$ or $\beta = 0$. Recursively define random variables $\overset{\bullet}{R}_K(\beta), \ldots, \overset{\bullet}{R}_0(\beta)$ by $\overset{\bullet}{R}_K(\beta) = \gamma\left(Y, \overline{L}(K), \overline{A}(K), \beta\right)$ and $\overset{\bullet}{R}_m(\beta) = \overset{\bullet}{r}_m\left(Y, \overline{L}(K), \overline{A}(K), \beta\right) = \gamma\left(\overset{\bullet}{R}_{m+1}(\beta), \overline{L}(m), \overline{A}(m), \beta\right)$ and set $\overset{\bullet}{R}(\beta) = \overset{\bullet}{r}\left(Y, \overline{L}(K), \overline{A}(K), \beta\right) \equiv \overset{\bullet}{R}_0(\beta)$. [Heuristically, $\overset{\bullet}{R}_m(\beta_0)$ is $Y_{\overline{A}(m-1),0}$ and $\overline{R}(\beta_0)$ is Y_0, where Y_0 is the outcome when treatment is always withheld. This is heuristic because in fact it is only the conditional distributions through time m that are guaranteed to be the same.] Also let $\overset{\bullet}{r}^{-1}\left(y, \overline{\ell}(K), \overline{a}(K), \beta\right)$ be the inverse of the function $\overset{\bullet}{r}$ with respect to its first argument. If $\gamma\left(y, \overline{\ell}(m), \overline{a}(m), \beta\right) = \gamma\left(y, \overline{a}(m), \beta\right)$ does not depend on $\overline{\ell}(m)$ for each m, we say that the SNDM model has no interaction.

Theorem 5.1: Under (4), a no interaction SNDM model is a stratified transformation model (STM), i.e., MSM 1c, with $R(\overline{a}, \beta) = \overset{\bullet}{r}(Y, \overline{a}, \beta)$ and $R(\beta) = \overset{\bullet}{R}(\beta)$. However, the converse is not true.

That is, a no-interaction SNDM is a MSM. The semiparametric information bound for β is greater if we correctly impose a no-interaction SNDM than if we only imposed the corresponding STM. An SNDM will be a MSM only if (as a fact of nature) there is no interaction. Theorem 5.1 indicates that a STM is the natural MSM analog of a SNDM in this case. If $\gamma\left(y,\overline{\ell}\left(m\right),\overline{a}\left(m\right)\right)$ depends on $\overline{\ell}\left(m\right)$, we must choose between analyzing the data under a SNDM versus a STM. To understand the advantages and disadvantages of each, we need some additional background. Define a regime $g=\left(g_0,\ldots,g_K\right)\in\mathcal{G}$ to be a collection of functions $g_m:\overline{\mathcal{L}}_m\rightarrow\mathcal{A}_m$. Define $g\left(\overline{\ell}\left(m\right)\right)=\left\{g_0\left(\bar{\ell}_0\right),\ldots,g_m\left(\overline{\ell}\left(m\right)\right)\right\}$. Let Y_g be the counterfactual value of Y if regime g were followed. If $g\left(\overline{\ell}_K\right)=\overline{a}_K\equiv\overline{a}$ does not depend on $\overline{\ell}_K$, then $Y_g=Y_{\overline{a}}$ and we say g is non-dynamic; otherwise, g is dynamic. Let $g\left[\overline{\ell}\left(k\right)\right]$ denote a realization of $\overline{A}\left(k\right)$. If we have sequential ignorability for regime g, i.e.,

$$Y_g\coprod A\left(t\right)\mid\overline{L}\left(t^{-}\right),\overline{A}\left(t^{-}\right)\ , \tag{22}$$

then, by Theorem 3.2 of Robins (1997), the law of Y_g is given by the G-computation algorithm formula

$$F_{Y_g}\left(y\mid\overline{\ell}\left(k\right),g\left[\overline{\ell}\left(k-1\right)\right]\right)=\iint F_Y\left(y\mid\overline{\ell}\left(K\right),g\left(\overline{\ell}\left(K\right)\right)\right)\prod_{m=k+1}^{K} dF\left[\ell\left(m\right)\mid\overline{\ell}\left(m-1\right),g\left(\overline{\ell}\left(m-1\right)\right)\right]\ . \tag{23}$$

We obtain $F_{Y_g}\left(y\right)$ from (23) by substituting in $k=-1$. Using the fact that, for continuous Y, Y is $\dot{r}^{-1}\left(\dot{R}\left(\beta_0\right),\overline{L}\left(K\right),\overline{A}\left(K\right),\beta_0\right)$, it can be shown that (23) implies

$$F_{Y_g}\left(y\right)=\iint I[\dot{r}^{-1}\left\{u,\overline{\ell}\left(K\right),g\left(\overline{\ell}\left(K\right)\right),\beta_0\right\}>y]\prod_{m=0}^{K} dF[\ell\left(m\right)\mid\overline{\ell}\left(m-1\right),g\left(\overline{\ell}\left(m-1\right)\right),\dot{R}\left(\beta_0\right)=u]dF_{\dot{R}\left(\beta_0\right)}\left(u\right)\ . \tag{24}$$

In many settings, the g-null hypothesis that

$$F_{Y_{g1}}\left(y\right)=F_{Y_{g2}}\left(y\right)\ \text{ for all } g1,g2\in\mathcal{G} \tag{25}$$

will be of interest. This is implied by the sharp null hypothesis of no treatment effect that $Y_{g1}=Y_{g2}$ with probability 1, i.e. no subject's outcome is influenced by the treatment history they choose. Robins (1986, 1997) proves the following.

Theorem 5.2: Given (22) for $g\in\mathcal{G}$, (25) holds $\Leftrightarrow$ (23) is the same for all $g\Leftrightarrow\gamma\left(y,\overline{\ell}_m,\overline{a}_m\right)=y\Leftrightarrow$

$$Y\coprod A\left(k\right)\mid\overline{L}\left(k\right),\overline{A}\left(k-1\right),k=0,\ldots,K\ . \tag{26}$$

5.2 Advantages of SNDMs with a Continuous Y

We are now ready to compare the advantages and disadvantages of SNDMs and MSMs for continuous Y. We begin by reviewing the advantages of SNDMs.

1. Although (26) implies $\beta_0=0$ for both a SNDM and a STM, only for a SNDM is (26) equivalent to $\beta_0=0$. What this means causally is the following. For a STM, the null hypothesis $\beta_0=0$ is equivalent to the hypothesis that the distribution of $Y_{\overline{a}}$ is the same for all non-dynamic regimes $\overline{a}$. This is a weaker hypothesis than the g-null hypothesis that says the distribution of Y_g is the same for all regimes, whether non-dynamic or dynamic. In most cases, it will be the latter null hypothesis (25) that will be of public health interest unless it were not possible to collect data on the covariates L_k which determine the treatment decisions for dynamic regimes.

2. If the $L(k)$ are discrete with only a moderate number of levels, then, even with $f\left[a(k) \mid \overline{\ell}(k-1), \overline{a}(k-1)\right]$ totally unrestricted, an asymptotically distribution-free g-null test of $\beta_0 = 0$ (and thus of (25)) exists for a SNDM but, because of the curse of dimensionality, not for a STM. Specifically, a non-parametric g-null test is equivalent to a test of independence of Y and $A(k)$ within strata defined jointly by $\overline{L}(k), \overline{A}(k-1)$ (Robins, 1997). Thus, even if $A(k)$ is continuous, a test of independence of $A(0)$ and Y within levels of $L(0)$ will be an asymptotic α - level test under (25). In contrast, a test of $\beta_0 = 0$ in a STM (without (25) additionally imposed) requires, by the Remark following Theorem 3.3a, that $\mathcal{W}$ can be consistently non-parametrically estimated which will not be possible due to the curse of dimensionality. (Note that to estimate $\mathcal{W}$, we must be able to consistently estimate the density of $A(k)$ given $\overline{L}(k)$ and $\overline{A}(k-1)$ for all k, which is not possible to do non-parametrically when the $A(m)$ are continuous.) In other words, the stronger hypothesis (25) that $\beta_0 = 0$ for a SNDM is easier to test non-parametrically than the weaker hypothesis that $\beta_0 = 0$ for a STM.

3. Henceforth, assume a correct model for $f\left[a(t) \mid \overline{\ell}(t^-), \overline{a}(t^-)\right]$ is available for all t. (We remind the reader that this would often be a false assumption.) Given a SNDM, with some difficulty the law of Y_g for dynamic g can be estimated using (24). In contrast, the law of Y_g for dynamic g is very hard to estimate given a STM. Specifically, given a SNDM, we estimate the law of Y_g as follows: *(i)* obtain an estimate $\widehat{\beta}$ by g-estimation (Robins, 1997), *(ii)* estimate $F_{\overset{\bullet}{R}(\beta_0)}(u)$ by the empirical law of the $\overset{\bullet}{R}_i\left(\widehat{\beta}\right)$ for $i = 1, \ldots, n$, *(iii)* specify and estimate a parametric model for $f\left[L(m) \mid \overline{L}(m-1), \overline{A}(m-1), \overset{\bullet}{R}\left(\widehat{\beta}\right)\right]$, *(iv)* and then evaluate the estimated version of the integral (24) by Monte carlo.

   In contrast, given a STM, we must, as discussed in Robins (1997, pg. 114; 1998a, Sec. 11) and Robins et al. (1999), specify a parametric model for $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right)$, where one can choose to define $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right)$ in either of two ways, leading to different parameterizations. Either $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right) \equiv \nu\left(y, \overline{\ell}(m), \overline{a}\right) - \nu\left(y, \left\{\overline{\ell}(m-1), \ell(m) = 0\right\}, \overline{a}\right)$ and $\nu\left(y, \overline{\ell}(m), \overline{a}\right)$ maps quantiles of $Y_{\overline{a}}$ given $\overline{\ell}_m, \overline{a}_{m-1}$ into quantiles of $Y_{\overline{a}}$ given $\overline{\ell}_{m-1}, \overline{a}_{m-1}$, or $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right)$ is defined to be the ratio of the hazard evaluated at y of $Y_{\overline{a}}$ given $\overline{\ell}_m, \overline{a}_{m-1}$ to the hazard at y of $Y_{\overline{a}}$ given $\overline{\ell}_{m-1}, \overline{a}_{m-1}, \ell(m) = 0$. Robins et al. (1999, Sec.8.7a) argue for the second option, since, in contrast to the first option, a parameterization in terms of hazard ratios is variation independent. As discussed in Robins (1997, pg. 114-116; 1998a, Sec. 11) and Robins et al. (1999), estimation of $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right)$ is a computational nightmare; indeed, fully parametric Bayesian or likelihood-based inference for a MSM is computationally extremely burdensome.

4. As discussed in Robins (1997, Sec. 9; 1998bc) and Robins et al. (1999), for SNDMs, it is easy to perform a sensitivity analysis in which the fundamental assumption (22) of ignorable treatment assignment is no longer imposed. For a STM such a sensitivity analysis is somewhat less straightforward and sensitivity analysis methods for MSMs are described in Robins et al. (1999) and Appendix 3 below.

5. A parameter β_0 of a SNDM, in contrast to that of a STM, can often still be consistently estimated if (22) is false but data are available on an instrumental variable. Specifically, suppose $A(t) = (A_1(t), A_2(t))$ with $A_1(t)$ recording a physician's prescribed treatment and $A_2(t)$ recording treatment actually received. We might suppose (22) is false, but $A_1(t) \coprod Y_g \mid \overline{L}(t^-), \overline{A}(t^-)$ is true if a predictor of Y_g and of $A_2(t)$ was not recorded in $\overline{L}(t^-)$. $A_1(t)$ is often then referred to as

an instrumental variable process, particularly when $A_1(t)$ has no direct causal effect, i.e., $Y_{\overline{a}} = Y_{\overline{a}_2}$ w.p.1. In this setting, the parameter of a STM is not identified but the parameter of a SNDM can still in general be consistently estimated by g-estimation (Robins, 1993; 1998b).

4. MSMs, in contrast to SNMs, cannot be used if there exists a value of ℓ_k, say $\ell_k = 0$, such that for all but one $a_k \in \mathcal{A}_k$, $f\left[a_k \mid \overline{\ell}_{k-1}, \ell_k = 0, \overline{a}_{k-1}\right] = 0$, since then the artificial censoring time $C^{\dagger}$ is zero with probability 1. An example would be a study of the effect of an occupational exposure on mortality with $\ell_k = 0$ if a subject is off work at time k, $\ell_k = 1$ otherwise, and subjects off work can only receive exposure level $a_k = 0$.

5.3 Advantages of MSMs with Continuous Y or with Failure Time Outcomes

1. Even in the presence of interaction [i.e., $\gamma\left(y, \overline{\ell}_m, \overline{a}_m\right)$ depends on $\overline{\ell}_m$], given a STM, the distribution of a non-dynamic counterfactual outcome $F_{Y_{\overline{a}}}(y)$ can be estimated by $n^{-1}\sum_i I\left\{r^{-1}\left[R_i\left(\widehat{\beta}\right), \overline{a}, \widehat{\beta}\right] > y\right\}$ without requiring either integration or modelling of the conditional law of $L(m)$. In contrast, as described in point 3 of Sec. 5.2 above, for a SNDM, both integration and modelling are required.

2. Any MSM that can be easily estimated when (8) holds (i.e., $\overline{A}$ is an ancillary process) can be easily estimated when (8) is false. For example, we can use the Cox proportional hazards MSM 2a for a continuous failure time outcome $T_{\overline{a}}$. In contrast, a structural nested Cox model would model the ratio of the conditional hazard given $\overline{\ell}_m, \overline{a}_m$ of $Y_{\overline{a}(m),0}$ to that of $Y_{\overline{a}(m-1),0}$ as a function of an unknown finite-dimensional parameter. Unfortunately, a structural nested Cox model does not admit any simple semiparametric estimators, and even complex estimators will fail due to the curse of dimensionality. Formally, the CODA information bound of Robins and Ritov (1997) for a structural nested Cox model is zero, even when $f\left(a_k \mid \overline{\ell}_k, \overline{a}_{k-1}\right)$ is completely known.

A possible hybrid approach is to impose a MSM model and then specify a model for $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right)$ is as follows. The g-null hypothesis (26) is true if and only if the distribution of $Y_{\overline{a}}$ given $V^{\dagger}$ is the same for all $\overline{a}$ (i.e., the parameter β_0 of our MSM is zero) and $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right)$ depends on $\overline{a}$ only through $\overline{a}_{m-1}$ (Robins,1997, Appendix 3). Thus, we impose a MSM model depending on a parameter β and an additional model for $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right)$ that depends on both the parameter β of the MSM model and another parameter $\psi = (\psi_1, \psi_2)$ in such a way that $\beta\psi_1 = 0$ if and only if $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right)$ depends on $\overline{a}$ only through $\overline{a}_{m-1}$. Specifically, $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right) / \nu^*\left(y, \overline{\ell}(m), (\overline{a}_{m-1}, a_m = 0, ..., a_K = 0)\right)$ depends on (β, ψ) only through the product $\beta\psi_1$ and $\nu^*(y, \overline{\ell}(m), (\overline{a}_{m-1}, a_m = 0, ..., a_K = 0))$ depends only on ψ_2. Thus ψ_1 is identified only if $\beta \neq 0$. Such a model can overcome objections 1 and 2 of Sec. 5.2 (but not objections 3-6) while retaining the advantages 1-2 of Sec. 5.3.

5.4 Structural Nested Mean Models (SNMMs)

We now turn to comparing MSMs and SNMs for discrete outcomes. Consider the set up of Sec. 5.1 but with Y discrete. For discrete outcomes, we define structural nested mean models. However, SNMMs are applicable to discrete and continuous outcomes.

Let $\gamma\left(\overline{\ell}(m), \overline{a}(m)\right) = E\left[Y_{\overline{a}(m),0} - Y_{\overline{a}(m-1),0} \mid \overline{\ell}(m), \overline{a}(m)\right]$. Let $\gamma^{\dagger}\left(\overline{\ell}(m), \overline{a}(m)\right) = ln\{E\left[Y_{\overline{a}(m),0} \mid \overline{\ell}(m), \overline{a}(m)\right] / E\left[Y_{\overline{a}(m-1),0} \mid \overline{\ell}(m), \overline{a}(m)\right]\}$. An additive structural nested mean model

(SNMM) specifies $\gamma\left(\overline{\ell}(m),\overline{a}(m)\right) = \gamma\left(\overline{\ell}(m),\overline{a}(m),\beta_0\right)$ with $\gamma\left(\overline{\ell}_m,\overline{a}_m,\beta\right)$ a known function satisfying $\gamma\left(\overline{\ell}_m,\overline{a}_m,\beta\right) = 0$ if $a_m = 0$ or $\beta = 0$. A multiplicative SNMM specifies $\gamma^{\dagger}\left(\overline{\ell}(m),\overline{a}(m)\right) = \gamma\left(\overline{\ell}(m),\overline{a}(m),\beta_0\right)$. The g-null mean hypothesis is the hypothesis

$$E\left[Y_{g1}\right] = E\left[Y_{g2}\right], g_1, g_2 \in \mathcal{G} \ . \tag{27}$$

Robins (1997) proves the following.

Theorem 5.2: Given (22), (27) holds if and only if $\gamma\left(\overline{\ell}(m),\overline{a}(m)\right) = 0 \Leftrightarrow \gamma^{\dagger}\left(\overline{\ell}(m),\overline{a}(m)\right) = 0 \Leftrightarrow E\left[Y \mid \overline{A}(k),\overline{L}(k)\right] = E\left[Y \mid \overline{A}(k-1),\overline{L}(k)\right], k = 0,\ldots,K$:

Advantages *(1) - (4)* and (6) of Sec. 5.2 of a SNM over a MSM for continuous Y also will hold (appropriately modified) for discrete Y when considering the g-null mean hypothesis or when estimating $E\left[Y_g\right]$.). Advantages *(1) - (2)* in Sec. 5.3 of a MSM over a SNDM for continuous outcome also hold in the discrete case.

An important advantage of MSMs over SNMs with Y dichotomous (or, more generally, when Y has finite support) is that neither an additive SNMM or multiplicative SNMM naturally imposes the fact that, for dichotomous Y, $E\left[Y_g\right] \in [0,1]$. In contrast, using the MSM model 1a with $g(\overline{a},\beta)$ a logistic function, the above restriction is naturally imposed. Analogously, in the setting of MSM model 1d, we can use standard marginal logistic models for the repeated measures outcomes $Y_{\overline{a}}(m)$. There exists logistic SNMMs that do impose that $E\left[Y_g\right] \in [0,1]$ (Robins et al., 1999). However, these logistic SNMMs are not very useful for semiparametric inference with high-dimensional data, since the CODA information bound for the parameter ψ of interest is zero even when $f\left(a_k \mid \overline{\ell}_k,\overline{a}_{k-1}\right)$ is known.

5.5 Direct Effect Models

In this section we show that some of the advantages of MSMs described in Secs. 5.3 and 5.4 are not retained in semiparametric models for the direct effect of a treatment a_1 when a second treatment a_2 is held fixed (set). In such a setting, both MSMs and direct effect SNMs (Robins, 1998a) have important limitations due to the curse of dimensionality if the functional form of the effect of the second treatment a_2 on the outcome is left completely unrestricted. Let $a(u) = (a_1(u), a_2(u))$ and, in a slight abuse of notation, set $\overline{a}(u) = (\overline{a}_1(u), \overline{a}_2(u))$ and $\overline{a} = (\overline{a}_1, \overline{a}_2)$. Continue to assume $V^{\dagger} = \emptyset$. Consider the following.

Model 3a – direct effect semiparametric regression: Consider the set-up of MSM 1b, with

$$\eta\left\{E\left[Y_{\overline{a}}\right]\right\} = g\left[\overline{a},\beta_0\right] + g^{\dagger}\left(\overline{a}_2\right)$$

where $g\left[\overline{a},\beta_0\right] = 0$ if $\overline{a}_1 \equiv 0$ and $g^{\dagger}(\cdot,\cdot)$ is unknown and unrestricted. Since, according to the model, $\eta\left\{E\left[Y_{\overline{a}_1,,\overline{a}_2}\right]\right\} - \eta\left\{E\left[Y_{\overline{a}_1\equiv 0,\overline{a}_2}\right]\right\} = g\left(\overline{a},\beta_0\right)$, it follows we are modelling the direct effect of treatment $\overline{a}_1$. Furthermore, the main effect of the second treatment $g^{\dagger}\left(\overline{a}_2\right) = \eta\left\{E\left[Y_{\overline{a}_1\equiv 0,\overline{a}_2}\right]\right\} - \eta\left\{E\left[Y_{\overline{a}_1\equiv 0,\overline{a}_2\equiv 0}\right]\right\}$ is completely unrestricted. Under sequential randomization assumption (4), the model for the observables O induced by MSM 3a is isomorphic to that induced by MSM 1b with $\overline{A}_2 \equiv \overline{A}_2(K)$ playing the role of $V^{\dagger}$. In particular, if $\eta(x) = x$ [or $ln(x)$], $\widehat{\beta}(h,\phi)$ will perform well in moderate size samples, provided $f\left[a(t) \mid \overline{L}(t^-),\overline{A}(t^-)\right]$ is known or can be parametrically modelled. However, as discussed in the final remark of Sec. 3.1, if $\eta(x) = ln\left[x/(1-x)\right]$, reasonable estimators of β_0 are unavailable because $\overline{A}_2(K)$ will be high-dimensional. Indeed, any choice of $\eta(x)$ that guarantees that $E\left[Y_{\overline{a}}\right] \in [0,1]$ will fail to provide reasonable estimators of β_0, negating the advantage of this MSM for dichotomous Y. This reflects the fact that the CODA information bound is zero, even when $f\left(a_k \mid \overline{\ell}_k,\overline{a}_{k-1}\right)$ is known.

Model 3b – direct effect semiparametric Cox proportional hazards model: Consider the set up of MSM 2b, with

$$\lambda_{T_{\overline{a}}}(t) = \lambda_{T_{\overline{a}_1=0,\overline{a}_2}}(t) \exp\left[r\left\{\overline{a}\left(t^-\right), t; \beta_0\right\}\right]$$

with $r\left(\overline{a}\left(t^-\right), t; \beta_0\right) = 0$ if $\overline{a}_1\left(t^-\right) = \mathbf{0}$. This is a model for the direct effect of treatment $\overline{a}_1$ on the hazard of T with the main effect of $\overline{a}_2$ left unrestricted. Given (4), MSM 3b induces a model for the observables isomorphic to that induced by MSM 2b. This implies that, as discussed in the remark in Sec. 3.1, due to the curse of dimensionality, it will not be possible to obtain reasonable estimators of β_0 negating advantage 2 of Sec. 5.3.

Model 3c – direct effect semiparametric time-dependent accelerated failure time model: Strikingly, the accelerated failure time MSM we now develop does not suffer from degradation due to the curse of dimensionality as did model 3b. Consider the model

$$\lambda_{R(\overline{a},\beta_0)}(t) = \lambda_{T_{\overline{a}_1=0,\overline{a}_2}}(t)$$

where $R(\overline{a}, \beta) = r\left(T_{\overline{a}}, \overline{a}, \beta\right)$ satisfies $r(t, \overline{a}, \beta) = t$ if $\overline{a}_1 = \mathbf{0}$ or $\beta = 0$. The model for the observables induced by MSM 3c is isomorphic to that induced by MSM 2c with $\overline{A}_2$ in the role of $V^\dagger$. Hence, the association model 3c can be used to estimate the direct effect of $\overline{a}_1$ on T with the main effect of $\overline{a}_2$ unrestricted. MSM 3c is the natural MSM associated with a structural nested failure time model (SNFTM) (Robins, 1993, App. 1; 1998) since a direct-effect SNFTM without interaction is a MSM 3c. In the presence of interaction, the MSM 3c retains advantage 1 of Sec. 5.3.

Appendix 1:

By arguments as in Robins et al. (1994), Theorem 3.1 and 3.3b are easy corollaries of Theorem 3.3a.

Sketch of Proof of Theorem 3.3a: For convenience, assume the L process and A process jump at times $0^-, 1^-, \ldots$ and $0, 1, \ldots$ respectively. Then by Theorem 3.2 of Robins (1997), Eq. (4) implies the G-computation algorithm formula

$$f_{\overline{Y}_{\overline{a}(k)}}\left[\overline{y}(k) \mid v^\dagger\right] = \tag{A1}$$
$$\int_{m=0}^{k} f\left[y(m), v(m) \mid \overline{y}(m-1), \overline{v}(m-1), \overline{a}(m-1), v^\dagger\right]_{m=1}^{k} d\mu\left[v(m)\right] d\mu\left(\overset{\bullet}{v}\right),$$

with $\overset{\bullet}{v} \equiv v(0) \backslash v^\dagger$. Thus, if for some $j < k$, the proposition $f\left[a(j) \mid \overline{a}(j-1), \overline{L}(j), v^\dagger\right] \neq 0$ *w.p.*1 given $V^\dagger$ is false, then (A1) is not identified. Hence, a MSM model places no (local) restrictions on $f\left[L(m) \mid \overline{L}(m-1), \overline{A}(m-1)\right]$ for $m > C^\dagger$. Hence, in semiparametric model *(ii)*, every function of O with mean zero given $O^\dagger$ is in the nuisance tangent space for the model. It follows that all members of $\Lambda^\perp$ in model *(ii)* depend on the data only through $O^\dagger$.

Because of our assumed knowledge of $\Lambda^{\perp,*}$ (the orthogonal complement to the nuisance tangent space under F^*), it is sufficient to show that $U \in \Lambda^\perp \Leftrightarrow U\mathcal{W} \in \Lambda^{\perp,*}$ when F^* is chosen such that $C^{\dagger,*}$ is equal to $C^\dagger$. This follows from the fact that $\overset{\circ}{\Lambda} = \overset{\circ}{\Lambda}{}^*$ where $\overset{\circ}{\Lambda} \equiv \Lambda \bigcap \left\{U_{tp}(\phi)\right\}^\perp \bigcap \left\{z\left(O^\dagger\right)\right\}$ and the fact that $E[UB] = E^*[U\mathcal{W}B]$ for any $B \in \overset{\circ}{\Lambda}$ by Lemma 3.1.

Appendix 2

Proof: In our model, Eq. (12) states that
$E[h\left(\overline{A}, V^\dagger\right) \varepsilon \mathcal{W}^{-1}\{h_{eff}\left(\overline{A}, V^\dagger\right) \varepsilon \mathcal{W}^{-1} - h_{eff}\left(\overline{A}, V^\dagger\right) \mathcal{W}^{-1} E\left[\varepsilon \mid \overline{A}, L(0)\right] +$

$E\left[h_{eff}\left(\overline{A},V^{\dagger}\right)\mathcal{W}^{-1}E\left[\varepsilon\mid\overline{A},L(0)\right]\mid L(0)\right]\}]=\kappa(h)\equiv E\left[h\left(\overline{A},V^{\dagger}\right)\mathcal{W}^{-1}\partial g\left(\overline{A},V^{\dagger},\beta_0\right)/\partial\beta\right]$. This can be rewritten as

$$\begin{aligned}E[h\left(\overline{A},V^{\dagger}\right)\{h_{eff}\left(\overline{A},V^{\dagger}\right)\mathcal{W}^{-2}var\left[\varepsilon\mid\overline{A},L(0)\right]+\mathcal{B}(h_{eff})\}]=\\ E\left[h\left(\overline{A},V^{\dagger}\right)\mathcal{W}^{-1}\partial g\left(\overline{A},V^{\dagger},\beta_0\right)/\partial\beta\right]\end{aligned}\tag{A2.1}$$

where $\mathcal{B}(h_{eff})=\varepsilon\mathcal{W}^{-1}E\left\{h_{eff}\left(\overline{A},V^{\dagger}\right)\mathcal{W}^{-1}E\left[\varepsilon\mid\overline{A},L(0)\right]\mid L(0)\right\}$. Now (A2.1) is true for all $h\left(\overline{A},V^{\dagger}\right)$ if and only if

$$\begin{aligned}h_{eff}\left(\overline{A},V^{\dagger}\right)E\left[\mathcal{W}^{-2}var\left[\varepsilon\mid\overline{A},L(0)\right]\mid\overline{A},V^{\dagger}\right]+\\ E\left[\mathcal{B}(h_{eff})\mid\overline{A},V^{\dagger}\right]=E\left[\mathcal{W}^{-1}\partial g\left(\overline{A},V^{\dagger},\beta_0\right)/\partial\beta\mid\overline{A},V^{\dagger}\right]\ .\end{aligned}\tag{A2.2}$$

To simplify (A2.2), note for any $q\left[\overline{A},L(0)\right]$, $E\left[q\left(\overline{A},L(0)\right)\mathcal{W}^{-1}\mid\overline{A},V^{\dagger}\right]=$
$f^*\left(\overline{A}\mid V^{\dagger}\right)\int q\left(\overline{A},L(0)\right)\left\{f\left(\overline{A}\mid L(0)\right)\right\}^{-1}\{f\left(\overline{A}\mid L(0),V^{\dagger}\right)f\left[L(0)\mid V^{\dagger}\right]/$
$f\left(\overline{A}\mid V^{\dagger}\right)\}d\mu\left(V^{\bullet}\right)=f^*\left(\overline{A}\mid V^{\dagger}\right)\left\{f\left(\overline{A}\mid V^{\dagger}\right)\right\}^{-1}\int q\left(\overline{A},L(0)\right)f\left(V^{\bullet}\mid V^{\dagger}\right)d\mu\left(V^{\bullet}\right)$. Thus, we have

$$E\left[\mathcal{W}^{-1}\mid\overline{A},V^{\dagger}\right]=f^*\left(\overline{A}\mid V^{\dagger}\right)/f\left(\overline{A}\mid V^{\dagger}\right)\ .\tag{*}$$

$$\begin{aligned}E\left[\mathcal{W}^{-2}var\left[\varepsilon\mid\overline{A},L(0)\right]\mid\overline{A},V^{\dagger}\right]=\\ f^*\left(\overline{A}\mid V^{\dagger}\right)\left\{f\left(\overline{A}\mid V^{\dagger}\right)\right\}^{-1}\int\mathcal{W}^{-1}var\left[\varepsilon\mid\overline{A},L(0)\right]f\left(V^{\bullet}\mid V^{\dagger}\right)d\mu\left(V^{\bullet}\right)\end{aligned}\tag{**}$$

and

$$\begin{aligned}E\left[\mathcal{B}(h_{eff})\mid\overline{A},V^{\dagger}\right]=\\ f^*\left(\overline{A}\mid V^{\dagger}\right)\left\{f\left(\overline{A}\mid V^{\dagger}\right)\right\}^{-1}\int E\left[\varepsilon\mid\overline{A},L(0)\right]f\left(V^{\bullet}\mid V^{\dagger}\right)d\mu\left(V^{\bullet}\right)\\ \left\{\int E\left[\varepsilon\mid\overline{a},L(0)\right]f^*\left(\overline{a}\mid V^{\dagger}\right)h_{eff}\left(\overline{a},V^{\dagger}\right)d\mu(\overline{a})\right\}=\\ \int h_{eff}\left(\overline{a},V^{\dagger}\right)f^*\left(\overline{a}\mid V^{\dagger}\right)d\mu(\overline{a})\,\omega\left(\overline{a},\overline{A},V^{\dagger}\right)\ .\end{aligned}\tag{***}$$

Substituting *, **, *** into (A2.2) proves the theorem.

Appendix 3: Sensitivity analysis for continuous and failure-time outcomes

Suppose rather than making the assumption (4) of sequential randomization, we instead assume for a model with a continuous outcome Y measured at end of follow-up at time $K+1^-$ (e.g., model 1c of Sec. 2.2) or a continuous failure-time outcome T (models 2a - 2c of Sec. 2.2) the existence of a known function $q_m\left(y,\overline{\ell}_m,\overline{a},a_m^*\right)$ such that for $Y_{\overline{a}}$ the continuous counterfactual variable measured at end of follow-up

$$pr\left[Y_{\overline{a}}<y\mid\overline{\ell}(m),\overline{a}(m-1),a^*(m)\right]=pr\left[q\left(Y_{\overline{a}},\overline{\ell}(m),\overline{a},a^*(m)\right)<y\mid\overline{\ell}(m),\overline{a}(m)\right]\tag{A3.1}$$

and, for $T_{\overline{a}}$ the counterfactual failure time variable,

$$pr\left[T_{\overline{a}}<y\mid\overline{\ell}(m),\overline{a}(m-1),a^*(m),T\geq m\right]=pr\left[q\left(T_{\overline{a}},\overline{\ell}(m),\overline{a},a^*(m)\right)<y\mid\overline{\ell}(m),\overline{a}(m),T\geq m\right]\ .\tag{A3.2}$$

The chosen conditional quantile - quantile function $q_m\left(y,\overline{\ell}_m,\overline{a},a_m^*\right)$ must satisfy

$$q_m\left(y,\overline{\ell}(m),\overline{a},a^*(m)\right)=y\ \ if\ \ a^*(m)=a(m)\tag{A3.3}$$

$$q_m\left(y, \overline{\ell}\,(m)\,, \overline{a}, a^*\,(m)\right) \textit{ is increasing in } y \tag{A3.4}$$

and, in the failure time case (A3.2),

$$q_m\left(y, \overline{\ell}\,(m)\,, \overline{a}, a^*\,(m)\right) > m \tag{A3.5a}$$

and writing $\mu = q_m\left(y, \overline{\ell}_m, \overline{a}, a^*_m\right)$, then

$$q_m\left(y, \overline{\ell}_m, \overline{a}, a^*_m\right) \textit{ is a function of } \overline{a} \textit{ only through } \overline{a}\left[\max\left(y, u\right)\right] \ . \tag{A3.5b}$$

This last restriction follows by consistency assumption (2). Note that the sequential randomization assumption (4) implies that $q_m\left(y, \overline{\ell}_m, \overline{a}, a^*_m\right) \equiv y$.

We now sketch how to construct a regular asymptotically linear (RAL) estimator of the parameter β_0 of a MSM such as model (1c) or model (2a) - (2c). Consider first the continuous outcome Y measured at end of follow-up. We shall replace each subject's observed outcome Y with J pseudo-Y's obtained by use of the following algorithm.

- Step 1: Do for $j = 1, \ldots, J$.
- Step 2: Do for $m = 0, \ldots, K$.
  - Draw $a^*_j\,(m)$ from $f\left[a\,(m) \mid \overline{L}\,(m)\,, \overline{A}\,(m-1)\right]$.
- Step 3: Set $Y_{(K+1)j} = Y$.
- Step 4: Do for $m = K, \ldots, 0$ $Y_{mj} = q_m\left(Y_{(m+1)j}, \overline{L}\,(m)\,, \overline{A}, a^*_j\,(m)\right)$.
- Create a new data set with $n \times J$ observations, $O_{ij} = \left(\overline{A}_i\,(K)\,, \overline{L}_i\,(K)\,, Y_{i0j}\right)$, $i = 1, \ldots, n, j = 1, \ldots, J$. Now for each observation O_{ij}, calculate $\widehat{D}_{sm}\,(\beta, a)$ and $D_{tp}\,(\phi)$ as described in the paragraph following Lemma 3.1, with Y_{i0j} in place of the actual data Y_i. Then let $\widehat{\beta}\,(h, \phi)$ solve $0 =_{i=1}^{n} {}_{j=1}^{J} \widehat{D}_{ij}\,(\beta, h, \phi)$ where, for each observation O_{ij}, $\widehat{D}\,(\beta, h, \phi) = \widehat{D}_{sm}\,(\beta, h) / W + D_{tp}\,(\phi)$.

Then it can be shown that, subject to regularity conditions, under the model characterized by (A3.1), an appropriate MSM, such as model 1c, and (11), $\widehat{\beta}\,(h, \phi)$ will be a RAL estimator of β_0.

The above algorithm can be modified so as to apply to a study with failure time outcomes under the assumption that the treatment process only jumps at times $0, 1, 2, 3, \ldots$ as follows.

To describe our algorithm for failure-time outcomes, we first discuss how to obtain a function $q_m\left(y, \overline{\ell}_m, \overline{a}, a^*_m\right)$ guaranteed to satisfy (A3.5a) and (A3.5b).

Define, for $m \leq [y-1]$ where $[x]$ is the greatest integer less than or equal to x, the function $q^*\left(x, \overline{\ell}\,(m)\,, \overline{a}\,[y]\,, a^*\,(m)\right)$ by

$$\begin{aligned} &pr\left[T_{(\overline{a}[y],0)} > u \mid \overline{\ell}\,(m)\,, \overline{a}\,(m-1)\,, a^*\,(m)\,, t \geq m, T_{(\overline{a}[y-1],0)} > [y]\right] = \\ &pr\left[q^*\left(T_{(\overline{a}[y-1],0)}, \overline{\ell}\,(m)\,, \overline{a}\,[y]\,, a^*\,(m)\right) > u \mid \overline{\ell}\,(m)\,, \overline{a}\,(m-1)\,, a^*\,(m)\,, T_{(\overline{a}[y-1],0)} > [y]\right] \ . \end{aligned} \tag{A3.6}$$

Then $q_m\left(y, \overline{\ell}\,(m)\,, \overline{a}, a^*\,(m)\right)$ is determined by $q^*_m\left(x, \overline{\ell}\,(m)\,, \overline{a}\,[y]\,, a^*\,(m)\right)$ and $q_m\left(y, \overline{\ell}\,(m)\,, \left(\overline{a}\,[y]\,, 0\right), a^*_m\right)$ through the following algorithm.

- Step 1: $p \leftarrow q_m\left(y, \overline{\ell}\,(m)\,, \left(\overline{a}\,[y]\,, 0\right), a^*_m\right)$.

- Step 2: Do for $k = 1, 2, \ldots$

  if $p < [y + k]$, stop and declare $q_m\left(y, \overline{\ell}(m), \overline{a}, a^*(m)\right) = p$.

  Otherwise, $p \leftarrow q_m^*\left(p, \overline{\ell}(m), \overline{a}[y + k], a^*(m)\right)$.

In conducting a sensitivity analysis, we choose $q_m\left(y, \overline{\ell}(m), (\overline{a}[y], 0), a^*(m)\right)$ and $q_m^*\left(x, \overline{\ell}(m), \overline{a}[y], a^*(m)\right)$ restricted only by the fact that the first function q_m is increasing in y, exceeds m and equals y if $a^*(m) = a(m)$ and that the second function q_m^* is increasing in y and exceeds $[y]$. Then we use the above algorithm to compute $q_m\left(y, \overline{\ell}(m), \overline{a}, a_m^*\right)$, which will then be guaranteed to satisfy (A3.5a) and (A3.5b).

We now describe how to construct a RAL estimator for the parameter of β_0 of a MSM failure time model such as model (2a). We assume we have on each of n subjects the data $\Delta = I(T = C), X = \min(T, C), \overline{A}(X), \overline{L}(X)$ where T is the failure time variable and C is the censoring variable. The algorithm goes as follows.

- Step 1: Do for $j = 1, \ldots, J$
- Step 2: Set $K = [X]$
- Step 3: For $s = K + 1, K + 2, \ldots$
  - Draw $a_j(s)$ from a chosen density $f^*\left[a(s) \mid \overline{A}_K, a_j(K + 1), \ldots, a_j(s - 1)\right]$
- Step 4: Set $X_{(K+1)j} = X$
- Step 5: Do for $m = K, K - 1, \ldots, 0$
  - Draw $a_j^*(m)$ from $f\left[a(m) \mid \overline{L}(m), \overline{A}(m - 1), T > m\right]$
  - Define $\overline{A}_j = \left(\overline{A}(K), a_j(K + 1), a_j(K + 2), \ldots\right)$ and set $X_{mj} = q\left(X_{(m+1)j}, \overline{L}(m), \overline{A}_j, a_j^*(m)\right)$
- Step 6: Create a new data set with $n \times J$ observations

$$O_{ij} = \left(\overline{A}_{ij}(X_{ioj}), X_{ioj}, \Delta_i\right) \ .$$

  We then fit the Cox model (2a) as we described previously in the paper but based on the $n \times j$ observations O_{ij}, with each observation on subject i associated with the same weight $\mathcal{W}_i$, where in calculating the numerator of $\mathcal{W}_i \equiv \mathcal{W}(X_i)$, we must use density f^* that was used in step 3 above. The resulting estimator will be consistent under the assumption that the hazard of censoring at time t given all the data only depends on the observed past.

Robins et al. (1999, Sec. 8.7b) discuss some potential problems with the sensitivity analysis methods discussed in this section due to the lack of a variation independent parameterization.

BIBLIOGRAPHY:

Chamberlain, G. 1987. Asymptotic Efficiency in Estimation with Conditional Moment Restrictions, *Journal of Econometrics,* 34: 305-324.

Chamberlain, G. 1988. Efficiency bounds for semiparametric regression. *Technical Report*, Department of Statistics, University of Wisconsin.

Gill, R.D., van der Laan, M.J., & Robins, J.M. 1997. Coarsening at random: characterizations, conjectures and counterexamples. *Proceedings of the First Seattle Symposium on Survival Analysis*, pp. 255-294.

Gill, R.D. and Robins, J.M. 1999. Causal inference from complex longitudinal data: The continuous case. Unpublished manuscript.

Heitjan, D.F., and Rubin, D.B. 1991. Ignorability and Coarse Data. *The Annals of Statistics*, 19: 2244-2253.

Holland, P. 1986. Statistics and Causal Inference. *Journal of the American Statistical Association,* 81: 945-961.

Lewis, D. 1973. Causation. *Journal of Philosophy,* 70: 556-567.

Liang, K-Y., and Zeger, S.L. 1986. Longitudinal Data Analysis Using Generalized Linear Model. *Biometrika*, 73: 13-22.

Lin, D.Y., Wei, L-J. 1989. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84: 1074-1078.

Newey, W.K. and McFadden, D. 1993. Estimation in large samples. **Handbook of Econometrics, Vol. 4**. Eds. McFadden, D., Engler, R. Amsterdam: North Holland.

Pearl J. 1995. Causal Diagrams for Empirical Research. *Biometrika*, 82: 669-688.

Robins J.M.1986. A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. *Mathematical Modeling*, 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. 1993. Analytic methods for HIV treatment and cofactor effects. **AIDS Epidemiology - Methodological Issues**. Eds. Ostrow DG; Kessler R. Plenum Publishing, New York. pp. 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. 1997. Causal inference from complex longitudinal data. In: **Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics (120)**. M. Berkane, Editor. NY: Springer Verlag. pp. 69-117.

Robins, J.M. 1998a. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. **Computation, Causation, and Discovery**. Eds. C. Glymour and G. Cooper. Cambridge, MA: The MIT Press. Forthcoming.

Robins, J.M. 1998b. Structural nested failure time models. **Survival Analysis**, P.K. Andersen and N. Keiding, Section Editors. **The Encyclopedia of Biostatistics**, P. Armitage and T. Colton, Editors. Chichester, UK: John Wiley & Sons. pp. 4372-4389.

Robins, J.M. 1998c. Correction for non-compliance in equivalence trials. *Statistics in Medicine*, 17: 269-302.

Robins, J.M. 1998d. Marginal structural models. *1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science,* pp. *1-10.*

Robins, J.M., and Greenland S. 1989. The probability of causation under a stochastic model for individual risk. *Biometrics,* 45: 1125-1138.

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

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

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

Robins, J.M., Rotnitzky, A., and Scharfstein, D.O. 1999. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: **Statistical Models in Epidemiology**. Halloran E., Editor. Springer-Verlag. Forthcoming.

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

Robins, J.M., and Tsiatis A.A. 1992. Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika,* 79: 311-319.

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

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. 1976. Inference and Missing Data. *Biometrika*, 63: 581-592.

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

Sasieni, P. 1992. Information bounds for the conditional hazard ratio in a nested family of regression models. *Journal of the Royal Statistical Society, Series B*, 54: 617-635.

P. Spirtes, C. Glymour, R. Scheines. 1993. **Causation, Prediction, and Search. Lecture Notes in Statistics 81**. New York: Springer-Verlag.

van der Vaart, A.W. 1991. On differentiable functionals. *Annals of Statistics,* 19: 178-204.