SENSITIVITY ANALYSIS FOR SELECTION BIAS AND UNMEASURED CONFOUNDING IN MISSING DATA AND CAUSAL INFERENCE MODELS

JAMES M. ROBINS*, ANDREA ROTNITZKY†, AND DANIEL O. SCHARFSTEIN‡

Table of Contents

*Departments of Epidemiology and Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115; robins@hsph.harvard.edu.

†Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115.

‡Department of Biostatistics, Johns Hopkins School of Hygiene and Public Health, Baltimore, MD 21205.

M. E. Halloran et al. (eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*

1. Introduction. In both observational and randomized studies, subjects commonly drop out of the study (i.e., become censored) before end of follow-up. If, conditional on the history of the observed data up to t, the hazard of dropping out of the study (i.e., censoring) at time t does not depend on the possibly unobserved data subsequent to t, we say drop-out is ignorable or explainable (Rubin, 1976). On the other hand, if the hazard of drop-out depends on the possibly unobserved future, we say drop-out is non-ignorable or, equivalently, that there is selection bias on unobservables. Neither the existence of selection bias on unobservables nor its magnitude is identifiable from the joint distribution of the observables. In view of this fact, we argue that the data analyst should conduct a "sensitivity analysis" to quantify how one's inference concerning an outcome of interest varies as a function of the magnitude of non-identifiable selection bias.

In Sections 2 and 3, we present a new class of non-parametric (just) identified (NPI) models that are useful for this purpose. These models are non-parametric (i.e., saturated) in the sense that each model in the class places no restrictions on the joint distribution of the observed data. Hence, each model in the class fits the observed data perfectly and cannot be rejected by any statistical test. Each model is (just) identified in the sense that the model identifies the distribution of the underlying full data (i.e., the distribution of the data that would have been observed in the absence of drop-out). Each NPI model in the class is indexed by a selection bias function that quantifies the magnitude of selection bias due to unobservables. Since each model is non-parametric, this selection bias function is not identified from the distribution of the observed data. However, we show that one can perform a sensitivity analysis that examines how inferences concerning functionals of the full data change as the non-identified selection bias function is varied over a plausible range. A nice feature of our approach is that, as discussed in Section 4, for each choice of the non-identified selection bias function, the full data functionals of interest can be estimated at $n^{\frac{1}{2}}$-rates using the modern theory of estimation in semiparametric models with missing data (Robins and Rotnitzky, 1992; Bickel et al., 1993; van der Vaart, 1991; Robins and Ritov, 1997). The results in Sec. 4 draw in part on as yet unpublished work by the authors (Scharfstein, Rotnitzky, and Robins, 1999).

In Sec. 5, we study in further detail a particular NPI model — the selection odds NPI model. This model is the unique NPI model that has both a "pattern mixture" (Little, 1993) and a selection model interpretation. Under this model, we derive an explicit formula, the selection bias g-computation algorithm formula, for the distribution of the full (i.e., complete) data. This formula generalizes the g-computation algorithm formula of Robins (1986) to the setting in which drop-out is non-ignorable.

There is a close connection between selection bias due to unobserved factors in follow-up studies with drop-out and selection bias due to unmeasured confounding factors in causal inference models. In Secs. 6 and 7,

we use this connection to generalize our NPI selection bias models to NPI causal inference models. Unfortunately, we show in Sec. 7.2 that there is major difficulty with trying to construct semiparametric estimators of the parameters of a NPI selection odds causal inference model. One solution to this difficulty is to give up the attempt to construct simple semiparametric estimators and, instead, use fully parametric likelihood-based inference. This approach is briefly discussed in the last remark of Sec. 8.6.1. A second and better approach is to develop alternative NPI causal inference models that simultaneously allow for unmeasured confounding and admit simple semiparametric estimators. This latter approach is considered in Sec. 8. In that section, we generalize both the structural nested models of Robins (1989, 1997b) and the marginal structural models of Robins (1998ab) to allow for selection bias due to unmeasured confounding.

In Secs. 9 and 10, we return to the subject of missing data models. The NPI missing data models discussed in Sections 2, 3 assume a monotone missing data pattern. In Section 10, we construct NPI models, the selection bias permutation missingness models, for non-monotone missing data with positive probability of complete observations. These models generalize the permutation missingness models of Robins (1997a).

In Section 11, we consider a Bayesian, as opposed to a sensitivity analysis, approach to summarizing our uncertainty.

Other work on NPI models with non-ignorable missing data includes the papers by Baker et al. (1992), Robins (1997a), Rotnitzky, Robins, and Scharfstein (1998), Zheng and Klein (1994, 1995), Slud and Rubenstein (1983), Klein and Moeschberger (1988), Nordheim (1984), Little (1994), and Moeschberger and Klein (1995). In contrast with our approach, except for Robins (1997a) and Rotnitzky, Robins, and Scharfstein (1998), the NPI models described in these papers do not allow for the incorporation of data on high-dimensional time-dependent covariate processes. The availability of such data has become very frequent in both longitudinal randomized and non-randomized studies.

Rosenbaum (1995) has published a considerable body of work on sensitivity analysis for selection bias and unmeasured confounding. However, Rosenbaum's approach differs from ours in that his approach is not based on a class of NPI models indexed by selection bias functions. Thus, in Rosenbaum's approach, causal contrasts of interest are not consistently estimated as a function of the non-identified strength of residual unmeasured confounding.

A large body of previous work, originating with Cornfield (1959), on sensitivity analysis in causal inference models has assumed the existence of an unmeasured confounder of U. In a sensitivity analysis, one varies the association of U with the outcome Y (within levels of treatment and measured confounders) and the association of U with the treatment (within levels of measured confounders) (Schlesselman, 1978; Rosenbaum and Rubin, 1983; Lin et al., 1998). In contrast, in our approach, we simply model

the association of the counterfactual outcome variable with the treatment (within levels of measured confounders). The advantage of our approach is that *(i)* there are many fewer sensitivity parameters to vary, and *(ii)* the (essentially impossible) decision as to whether to view U as univariate or multivariate, continuous or discrete is done away with. A link between the two approaches is that the counterfactual variable can be considered the ultimate unmeasured confounder U. This reflects the fact that, given the counterfactual and treatment, other unmeasured covariates U fail to predict the observed outcome (and thus are superfluous and can be dispensed with), since the observed outcome variable is a deterministic function of the treatment and the counterfactual outcome.

In our opinion, the unmeasured confounder U approach should be preferred to our counterfactual approach only in circumstances, where *(i)* U represents a known confounder (e.g., cigarette smoking) that for logistical reasons was not measured in a particular study, and furthermore, *(ii)* there exists reasonable historical knowledge about the magnitude of association of U with both the outcome (conditional on treatment and measured confounders) and the treatment (conditional on measured confounders). In contrast, when U is to represent all possible unmeasured factors, we believe that it is substantively easier for subject-matter experts to give their opinions about the plausible magnitude of the association of the counterfactual outcome with treatment than about the question of whether any unmeasured confounders U are continuous or discrete, single or multidimensional, and the associations of such confounders with treatment and the outcome.

In the setting of a single time-independent treatment, our counterfactual approach leads to extremely simple computations that can be carried out with standard software. Specifically, as described in Remark 8.14 of Sec. 8.5, when the outcome Y is dichotomous, it is possible to use a standard logistic regression program equipped with an offset option that allows the analyst to fix coefficients of certain regressors to known values. In contrast, as discussed by Lin et al. (1998), there can be formidable computational difficulties associated with the approach based on positing an unmeasured covariate U. Even in the presence of time-varying treatments and covariates, the counterfactual approach to sensitivity analysis can remain computationally simple provided one uses the semiparametric estimation methods described in Sec. 8.

2. Identification in monotone missing data problems.

2.1. Single occasion. Consider a study designed such that a vector or scalar variable L_i is to be measured on each of n units $i, i = 1, \ldots, n$. The entire vector L_i is missing on a subset of the units. We therefore observe n i.i.d. copies of

$$O = (\Delta, \Delta L) \tag{2.1}$$

where $\Delta = 1$ if L is observed and $\Delta = 0$ otherwise. Throughout we refer to L_i as the *full data* on unit i. Thus, under our assumptions, (L_i, Δ_i) and $O_i, i = 1, \ldots, n$ are n i.i.d. copies of random variables (L, Δ) and O with cumulative distribution functions denoted by $F_{L,\Delta}$ and F_O respectively. In what follows, F_L is used to denote the cumulative distribution function of L. Here and henceforth we suppress the i subscript denoting unit. Little and Rubin (1987) refer to the product variable ΔL as L_{obs}. We assume that

$$pr(\Delta = 1) \neq 0. \tag{2.2}$$

The simple paradigmatic theorem underlying many of the results of this paper is the following.

In the following theorem and until Sec. 8, $\Phi(x)$ will denote a known, continuous, monotone increasing distribution function [i.e., $\Phi(x)$ is strictly increasing in x and $\Phi(x) \to 0$ as $x \to -\infty$ and $\Phi(x) \to 1$ as $x \to \infty$].

THEOREM 2.1. *Given (i) a law F_O of O satisfying (2.2), (ii) a continuous, monotone increasing, distribution function $\Phi(x)$, and (iii) a function $q(L)$, there exists a unique joint law $F_{L,\Delta}$ for (L, Δ) with marginal F_O for O satisfying*

$$pr[\Delta = 1 \mid L] = \Phi[h + q(L)] \tag{2.3}$$

for some $h \in (-\infty, \infty]$. Specifically, the constant h is the unique solution to

$$E_O[\Delta / \Phi\{h + q(L)\}] = 1 \tag{2.4}$$

and the marginal F_L of L under $F_{L,\Delta}$ is

$$F_L(\ell) = E_O[\Delta I(L \leq \ell) / \Phi\{h + q(L)\}] \tag{2.5}$$

where $E_O[\cdot]$ denotes expectation w.r.t. F_O, and, for multivariate L and ℓ, $L \leq \ell$ means that each component of L is less than or equal to the corresponding component of ℓ.

REMARK 2.1. If $q(L)$ is a constant not depending on L, then (2.3) says that, under $F_{L,\Delta}$, the data are missing at random in the sense defined by Rubin (1976). A choice $q(L)$ that is a non-constant function of L corresponds to Rubin's (1976) definition of a non-ignorable non-response process. Note that in the preceding Theorem, h can take the value $+\infty$. Indeed, h is equal to $+\infty$ if and only if $pr(\Delta = 1) = 1$. Eq. (2.2) guarantees h will always exceed $-\infty$.

REMARK 2.2. **Semiparametric model a:** Consider the semiparametric model **a** for the law of (Δ, L) defined by the following conditions:

$$F_L \text{ is unrestricted,} \tag{2.6}$$

$pr(\Delta = 1 \mid L) = \Phi\{h + q(L)\}$ where $q(\cdot)$ is a fixed and known function, $\Phi(\cdot)$ is a known, strictly increasing, distribution function, and h is unknown and ranges over $(-\infty, +\infty]$. By Theorem 2.1, this model places no restrictions on the law F_O of the observables and it is therefore non-parametric for the law F_O of the observed data. That is, Theorem 2.1 establishes that every F_O is the marginal of a law $F_{L,\Delta}$ allowed by the model, and thus for each choice of $q(L)$ the model fits the data perfectly and cannot be rejected by any statistical test. Second, the model is identified in the sense that the unknowns F_L and h in the model, and thus the joint distribution $F_{L,\Delta}$ are uniquely determined by the law F_O of the observed data. We refer to model **a** as a non-parametric (just) identified (NPI) model for (Δ, L). Robins (1997b) referred to NPI models as non-parametric saturated.

The NPI model **a** is useful for conducting sensitivity analysis as the following example illustrates.

REMARK 2.3. In Theorem 2.1 and Remark 2.2, we implicitly assume that the law F_O lay in some large set $\mathcal{F}$ of possible laws for O. In Theorem 2.1, as well as in similar remarks in Theorems to follow, we assume that the non-ignorable selection bias function q belongs to a set of functions $\mathcal{Q}$ satisfying the restriction that for all $F_O \in \mathcal{F}$ and $q \in \mathcal{Q}$, all expectations and integrals occurring in the statement of the theorem are finite. For example, if $\mathcal{F} = \{F_O; F_{L|\Delta=1}$ *is absolutely continuous with respect to Lebesgue measure*$\}$, i.e., L is a continuous random variable with potentially unbounded support, then $\mathcal{Q} = \{q(\ell); |q(\ell)|$ *is bounded by a constant* $c\}$ is a suitable choice for $\mathcal{Q}$. In contrast, if $\mathcal{F} = \{F_O; F_{L|\Delta=1}$ *has support on* $\{0,1\}\}$ (i.e., L is dichotomous), the set $\mathcal{Q}$ would be unrestricted. If we included laws in $\mathcal{F}$ for which the aforementioned integrals and expectations did not exist then model **a** would not be a nonparametric model since there would exist laws F_O in $\mathcal{F}$ which would not be the marginal of any law for (L, Δ) satisfying (2.3).

REMARK 2.4. Consider the following example.

EXAMPLE 2.1. Consider the special case of semiparametric model **a** in which $\Phi(x) = e^x/(1+e^x)$, L is scalar, and $q(L) = \alpha_1 L$. Then α_1 "quantifies" the magnitude of selection bias on a logistic scale with $\alpha_1 = 0$ denoting no selection bias, i.e., missing at random data. Now suppose the parameter of interest is $F_L(10)$, the marginal probability that L is less than 10 (were there no missing data). Then, given a law F_O for the observed data, by (2.4) and (2.5) we obtain $F_L(10)$ as a function of α_1 as shown in the schematic graph below. Figure 1 demonstrates that our inferences about $F_L(10)$ based on F_O depend on our choice of α_1. Note that α_1 is not identified by F_O since, by Theorem 2.1, F_O is perfectly compatible with any choice of α_1. Thus, in Figure 1, we have specified a range of values for α_1 and used these together with F_O to identify $F_L(10)$ as a function of the chosen α_1.

In practice, the law F_O of O is unknown but can be estimated at the usual $n^{\frac{1}{2}}$-rate by the empirical law F_n of the data that puts mass

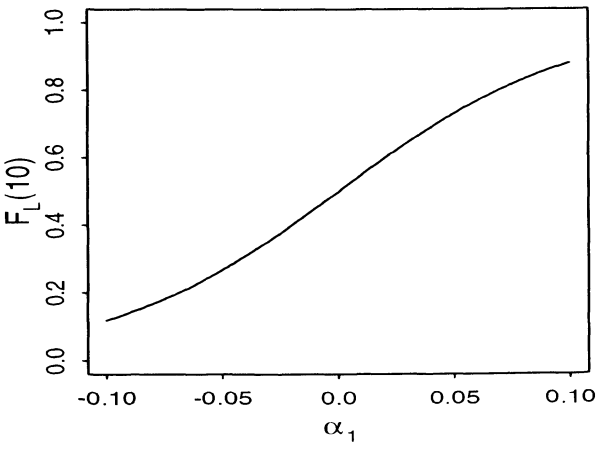

FIG. 1. $F_L(10)$ *as a function of* α_1.

$1/n$ on each of the n observations $(\Delta_i, \Delta_i L_i)$. Thus, in practice we would replace Figure 1 by Figure 2 where the solid line is the estimate $\widehat{F}_L(10)$ of $F_L(10)$ computed under F_n and the vertical bars between the dashed curves represent a 95 percent confidence interval for $F_L(10)$. $\widehat{F}_L(10)$ is obtained by the empirical versions of (2.4) and (2.5) in which $E_O[\cdot]$ is replaced by $\widetilde{E}_n[\cdot]$ where, for any random variable, $H_i, \widetilde{E}_n[H] = n^{-1}\sum_{i=1}^n H_i$. We delay to Section 4 describing how confidence intervals for $F_L(10)$ can be computed.

Since Theorem 2.1 implies that there will never be any data evidence that can determine either the magnitude of α_1 or that the function $q(\ell)$ is a linear function of ℓ, it follows that we might wish to repeat the preceding sensitivity analysis for other functions $q(\ell)$ such as $\alpha_1 e^{\ell}$. Note that this substantive meaning of the magnitude of α_1 depends on whether we choose $q(L)$ to be linear in L or exponential in L.

Two functions $q_1(L)$ and $q_2(L)$ that differ by a constant K result in the same semiparametric model (a) for the law of (Δ, L) because $\Phi(h + q_1(L)) = \Phi(h^* + q_2(L))$ where $h^* = h + K$ and the constant h is allowed to vary over the entire extended interval $(-\infty, \infty]$. It is therefore desirable to restrict the class of functions $q(L)$ over which a sensitivity analysis is to be conducted to a class of functions where different functions yield different models for the law of (Δ, L). One way to accomplish this is to restrict attention to functions $q(L)$ such that $q(0) = 0$. With this restriction, the

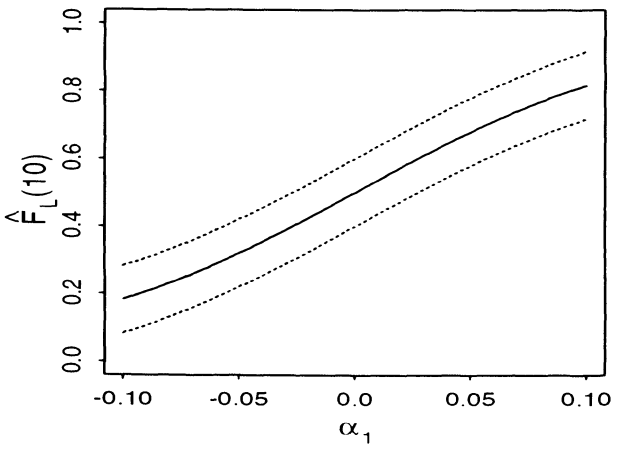

FIG. 2. $\widehat{F}_L(10)$ *as a function of* α_1, *with 95% confidence intervals.*

function $q(L) = 0$ is the unique function that corresponds to the data being missing at random.

Bounds – It would be advantageous to use, in a sensitivity analysis, parameterized functions $q(L) \equiv q(L;\alpha)$ such that in the limit as the parameter $\alpha \to -\infty$ and $\alpha \to \infty$, the implied laws for $F_L(\ell)$ based on the selection bias function $q(L;\alpha)$ converges to the bounds on $F_L(\ell)$ imposed by the law F_O of O. See Sec. 7 of Scharfstein, Rotnitzky, and Robins (1999) for additional discussion.

The reader may feel discouraged that we have only obtained a "sensitivity analysis" for the parameter of interest $F_L(10)$ and that we have been unable to jointly identify this parameter and the selection bias parameter α_1. However, we feel quite the opposite. Since the parameter α_1 represents the magnitude of selection bias due to unmeasured factors, it would not be desirable or scientifically reasonable for α_1 to be identified from the data in the absence of further knowledge of these factors. We would hope it would *not* be identifiable from the data without further well-supported substantive knowledge. Our semi-parametric model a for (L, Δ) operationalizes this desiderata; we cannot identify the magnitude of selection bias, but we can identify the law F_L as a function of the selection bias parameter.

To clarify this point, consider the parametric submodel of our semi-parametric model **a** in which we assume that L is normally distributed with unknown mean μ and variance σ^2, and again $\Phi(x)$ is logistic, $q(L;\alpha_1) =$

$\alpha_1 L$, and the "intercept" h in (2.3) is unknown. This parametric model is a special case of the "selection model" proposed by Heckman (1976). The parametric likelihood function for this model is

$$f\left(\Delta, \Delta L; \mu, \sigma^2, h, \alpha_1\right) = \left[\Phi\left(h + \alpha_1 L\right) f\left(L; \mu, \sigma^2\right)\right]^{\Delta} \times \left[\int_{-\infty}^{\infty} \left\{1 - \Phi\left(h + \alpha_1 L\right)\right\} f\left(L; \mu, \sigma^2\right) dL\right]^{1-\Delta}. \tag{2.7}$$

Even if α_1 is unknown, $\left(\mu, \sigma^2, h, \alpha_1\right)$ are all identified when $f\left(L; \mu, \sigma^2\right)$ is a normal density function (Heckman, 1976; Rotnitzky and Robins, 1997). However, our ability to identify both the law of L and the selection parameter α_1 comes solely from the assumption of normality. For example, as noted by Little and Rubin (1987), we can determine whether $\alpha_1 = 0$ (i.e., whether the data are MAR) by checking whether the observed conditional distribution of L given $\Delta = 1$ is skew. This follows because if $\alpha_1 = 0$ and L is normal, $f\left(L \mid \Delta = 1\right)$ is normal and thus is not skew; however, for all $\alpha_1 \neq 0$, $f\left(L \mid \Delta = 1\right)$ is skew. Thus, unless we know *a priori* with certainty that $f\left(L\right)$ is not skew, we should not use such a parametric model to test for selection bias because identification of α_1 comes entirely from the assumed distributional shape. As noted by Little and Rubin (1987), it is rare, if ever, that we would have such prior knowledge. In contrast, the likelihood in our semi-parametric model **a** in which $F_L\left(\cdot\right)$ is left unspecified is except for being dominated by Lesbesgue measure,

$$f\left(\Delta, \Delta L; \theta, h, \alpha_1\right) = \left[\Phi\left(h + \alpha_1 L\right) f\left(L; \theta\right)\right]^{\Delta} \times \left[\int_{-\infty}^{\infty} \left\{1 - \Phi\left(h + \alpha_1 L\right)\right\} f\left(L; \theta\right) dL\right]^{1-\Delta} \tag{2.8}$$

where θ is an infinite dimensional parameter indexing the laws of L that are dominated by a Lesbesgue measure and $\left(\theta, h, \alpha_1\right)$ are not jointly identified. However, given α_1 known, the NPMLE of h and $F_L\left(\ell\right)$ is given precisely by (2.4) and (2.5) with E_O replaced by $\widetilde{E}_n$ (where, as is usual in defining the NPMLE, when maximizing (2.8) we allow $f\left(L; \theta\right)$ to be any distribution, including discrete distributions). The NPMLE of F_L is a discrete distribution which jumps only at the observed values of L among subjects with $\Delta = 1$.

REMARK 2.5. Although the NPMLE can be used for sensitivity analysis in the simple "toy" selection model of this subsection, it will fail in the more realistic models discussed in the next two subsections due to the curse of dimensionality (Robins and Ritov, 1997). Estimation of more realistic semiparametric models uses estimating equations derived from the theory of semiparametric models (Bickel et al., 1993; Rotnitzky and Robins, 1997; Rotnitzky, Robins, and Scharfstein, 1998).

2.2. Longitudinal monotone missing data in discrete time. We now generalize our results to a longitudinal study in which the full data is $\overline{L}_{K+1} = (L_0, \ldots, L_{K+1})$ where L_k is a scalar or vector variable measured at occasion $k, k = 0, \ldots, K+1$, and we adopt the notation that, for any time-dependent variable Z_k, $\overline{Z}_k = (Z_0, Z_1, \ldots, Z_k)$ is the history of the variable through time k. By convention, we set $Z_{-1} = \overline{Z}_{-1}$ equal to zero with probability 1. We assume some subjects drop out of the study so that the observed data is $O = \left(C, \overline{L}_C\right)$ where the censoring time C denotes the last occasion on which L_k is observed. We assume that missingness is monotone so that if L_k is not observed, then L_{k+1} is also not observed. Note $C = K+1$ for a subject whose full data $\bar{L}_{K+1}$ are observed. We use $F_{C,L}$ to denote the joint distribution of C and L and $\lambda_k(L)$ to denote the conditional discretized hazard of censoring at time k given L, i.e., $\lambda_k(L) = pr(C = k \mid C \geq k, L)$. For notational convenience, we will often use Λ_k to denote the random variable $\lambda_k(L)$. Note we are not using Λ_k to denote a cumulative hazard function. Then let $\Delta = I(C = K+1)$ be the indicator for observing full data and denote $\overline{L}_{K+1}$ by L. We assume that (2.2) holds. Then in analogy to Theorem 2.1, we have the following Theorem 2.2. Its proof is given in Rotnitzky, Robins, and Scharfstein (1998). Theorem 2.1 above is a special case of Theorem 2.2.

THEOREM 2.2. *Given (i) a law F_O of O satisfying $pr[C \neq k \mid C \geq k, \overline{L}_k] \geq \sigma > 0$ w.p.1, $k = 0, \ldots, K$, (ii) a continuous, monotone increasing, distribution function $\Phi(x)$, and (iii) known functions $q_k(L), k = 0, \ldots, K$, there exists a unique joint law $F_{C,L}$ for (C, L) with marginal F_O for O satisfying the condition that, for $k = 0, \ldots, K$, one minus the discrete hazard $\Lambda_k = \lambda_k(L)$ of becoming censored at k given L is*

$$1 - \Lambda_k \equiv pr[C \neq k \mid C \geq k, L] = \Phi\left[h_k\left(\overline{L}_k\right) + q_k(L)\right] \tag{2.9}$$

for some functions $h_k\left(\overline{\ell}_k\right)$ taking values in $(-\infty, \infty]$. Specifically, $h_K(\cdot)$ is the unique solution to

$$E_O\left\{\Delta / \Phi\left[h_K\left(\overline{L}_K\right) + q_K(L)\right] \mid \overline{L}_K, C \geq K\right\} = 1\ ; \tag{2.10}$$

$h_m(\cdot)$, $m = K-1, \ldots, 0$ is recursively obtained as the unique solution to

$$E_O\left[\Delta / \left\{\prod_{j=m+1}^{K} (1 - \Lambda_j)\, \Phi\left[h_m\left(\overline{L}_m\right) + q_m(L)\right]\right\} \mid \overline{L}_m, C \geq m\right] = 1, \tag{2.11}$$

(where the $\Lambda_j, j = m+1, \ldots, K$ have been previously obtained), and the marginal F_L of L is

$$F_L(\ell) = E_O\left[\Delta I(L \leq \ell) / \prod_{j=0}^{K} (1 - \Lambda_j)\right]. \tag{2.12}$$

REMARK 2.6. **Semiparametric model a**: In analogy to Section 2.1, consider the semiparametric model **a** for the law of (C, L) defined by the following conditions:

F_L *is unrestricted*

$1 - \lambda_k(L) = \Phi\{h_k(\overline{L}_k) + q_k(L)\}$ *where* $q_k(\cdot)$ *is an arbitrary fixed and known function* $k = 0, \ldots, K$,

$\Phi(\cdot)$ *is a known, strictly increasing, distribution function, and* $h_k(\cdot)$ *is unknown and ranges over* $[-\infty, \infty]$.

Theorem 2.2 implies that model **a** places no restrictions on the law F_O of O and, furthermore, that the unknowns F_L and $h_k(\cdot)$ are identified. That is, model **a** is non-parametric (just) identified. In particular, the functions $q_k(\cdot)$ are not amendable to empirical verification, i.e., cannot be rejected by any statistical test, because all choices of $q_k(\cdot)$ are compatible with the observed data. As in Section 2.1, in order to avoid the possibility that different functions $q_k(\cdot)$ yield the same semiparametric model for $F_{C,L}$, we shall restrict attention to functions $q_k(\cdot)$ such that $q_k(L) = 0$ if $\underline{L}_{k+1} \equiv (L_{k+1}, \ldots, L_{K+1}) = 0$. Note if the restriction on the law F_O in *(i)* of Theorem (2.2) did not hold then, in general, identification will fail.

REMARK 2.7. If all the L_m are discrete with only a few levels, the NPMLEs of $h_k(\bar{\ell}_k)$ and $F_L(\ell)$ are obtained by substituting the population expectation E_O by its sample version in (2.10)–(2.12). In general, however, if the L_m has continuous components or many discrete components (L_m may be a random vector), then, due to the curse of dimensionality, the NPMLE will be undefined. As discussed in Section 4, in practice, inference in such settings requires placing additional restrictions on the functions $h_k(\cdot)$. One such approach would be to impose only smoothness conditions on the functions $h_k(\cdot)$ and to conduct inference about any functional $\beta(F_L)$ of interest (e.g., the mean of L_{K+1}) using smoothing methods to estimate the unknowns $h_k(\cdot)$. However, when the L_m have multiple continuous components, the use of multivariate smoothing techniques for estimating $h_k(\cdot)$ requires impractically large samples. Thus, in practice, more restrictive models are required on the functions $h_k(\cdot)$. Formally, we assume that $h_k(\bar{\ell}_k)$ lies in a class of functions $h_k(\bar{\ell}_k; \gamma)$ indexed by a parameter $\gamma \in \Gamma$ where the parameter space Γ may be finite dimensional or infinite dimensional. Therefore, let the semiparametric model $\mathbf{a}_r$ be semiparametric model **a** modified in that $h_k(\bar{\ell}_k)$ is subject to the restriction that it lies in $\mathcal{H}_k \equiv \{h_k(\bar{\ell}_k; \gamma); \gamma \in \Gamma\}$. When $\mathcal{H}_k$ does not contain all possible functions $h_k(\bar{\ell}_k)$, Theorem 2.2 implies that then there will exist laws F_O that are not the marginal of any law of (L, C) that lies in model $\mathbf{a}_r$ (i.e., $\mathbf{a}_r$ is not a non-parametric model for F_O). Thus, in principle, model $\mathbf{a}_r$ can be subjected to an empirical goodness-of-fit test. Our informal suggestion for conducting sensitivity analysis in this setting is first to *(i)* choose the

model for the $h_k(\bar{\ell}_k)$ large enough that nearly any goodness-of-fit test will have little power to reject the model $\mathbf{a}_\tau$ but *(ii)* choose it small enough so that the estimators described in Sec. 4 below of functionals $\beta(F_L)$ of interest have a nearly normal sampling distribution with the variance small enough to be substantively useful to subject matter experts. It is not clear that both criteria *(i)* and *(ii)* can always be met. Clearly the choice of the size of the model will depend on the size of the data set, the complexity of the functional $\beta(F_L)$, and the precision required by the subject matter experts. Furthermore, since different models $\mathbf{a}_\tau$ associated with different model choices for the $h_k(\bar{\ell}_k)$ cannot be easily distinguished by goodness-of-fit tests when the advice in *(i)* is followed, we suggest using a large number of different models for $h_k(\bar{\ell}_k)$. Then, for each model for the $h_k(\bar{\ell}_k)$, estimate F_L under many choices for the selection bias functions $q_k(L)$ (still treated as known in the analysis) so that sensitivity both to the choice of the $q_k(L)$ and to the choice of model for $h_k(\bar{\ell}_k)$ can be assessed.

2.3. Longitudinal monotone missing data in continuous time. In this section, we generalize the results of Section 2.2 to allow for drop-out (censoring) in continuous time. The full data are n i.i.d. copies of a, possibly multivariate, continuous time stopped stochastic process, $L \equiv \overline{L}(\tau)$, where for any stochastic process $Z(u)$, $\overline{Z}(t) = \{Z(u); 0 \leq u \leq t\}$ is the history of the process up to t, and the stopping time τ is the (possibly random) administrative end-of-follow-up time. Since τ is the administrative end-of-follow-up time, it is assumed known for each subject at time 0 and thus τ is a component of $L(0)$.

The observed data are now n i.i.d. copies of $O = \left(C, \overline{L}(C)\right)$ where C is time to drop-out and, by convention, $C \equiv \tau$ if the subject is uncensored (does not drop out) by administrative end-to-follow-up. In analogy to Sec. 2, let $\Delta = I(C = \tau)$ be the indicator that the full data $L = \overline{L}(\tau)$ are observed and we continue to assume (2.2) holds. Furthermore, we let F_O and $F_{C,L}$ denote the CDF of O and (C, L) respectively, and we use $\lambda(u \mid A)$ to denote the conditional hazard of C at u given A, that is, $\lambda(u \mid A) = \lim_{t \to 0} t^{-1} pr(u < C < u + t \mid C \geq u, A)$. Further, we write $\Lambda(u)$ to denote the random variable $\lambda(u \mid L)$. Note $\Lambda(u)$ is not a cumulative hazard function. We then have the following theorem whose proof is given in Appendix 1 of Scharfstein, Rotnitzky, and Robins (1999) in the special case in which the process $\overline{L}(u)$ only jumps at a finite number of fixed non-random times. Richard Gill has proved this theorem for more general processes. His proof will be reported elsewhere.

THEOREM 2.3. *Given (i) a law F_O of $O = \left(C, \overline{L}(C)\right)$ such that the hazard of C at u given $\overline{L}(u)$ is bounded by a constant c w.p.1 and (ii) a function $q(u, L), u \in [0, \tau)$, there exists a unique joint law $F_{C,L}$ with marginal F_O satisfying, for some function $h\left(u, \overline{L}(u)\right)$*

$$\lambda(u \mid L) = exp\left[h\left(u, \overline{L}(u)\right) + q(u, L)\right] . \tag{2.13}$$

Specifically, $h\left(u, \overline{L}(u)\right)$ *is the unique function taking values in* $(-\infty, \infty]$ *satisfying*

$$E_O\left[\Delta / \exp\left[-\int_u^\tau \lambda(x \mid L)\, dx\right] \mid \overline{L}(u), C \geq u\right] = 1 \tag{2.14}$$

and the marginal F_L *of* L *is*

$$F_L(\ell) = E_O\left[\Delta I(L \leq \ell)/S\right] \tag{2.15}$$

where

$$S = \exp\left[-\int_0^\tau \lambda(x \mid L)\, dx\right] \tag{2.16}$$

and $L \leq \ell$ *means each component of the vector* $L(u)$ *is less than or equal to the corresponding component* $\ell(u)$ *for all* $u \in [0, \tau]$.

REMARK 2.8. **Semiparametric model a:** As in Section 2.2, consider the semiparametric model a for the law of (C, L) defined by the conditions *(i)* F_L is unrestricted and *(ii)* $\lambda(u \mid L)$ is given by (2.13) with the selection bias function $q(u, L)$ known and $h\left(u, \overline{L}(u)\right)$ completely unrestricted.

REMARK 2.9. It follows from Theorem 2.3 that model **a** is nonparametric (just) identified. In analogy to Section 2.2, we will restrict attention to a class of functions $q(u, L)$ for which each member results in a different semiparametric model for the law of (C, L). Specifically, we will consider functions $q(u, L)$ satisfying $q(u, L) = 0$ if $\underline{L}(u) \equiv \{L(t); u < t \leq \tau\}$ is equal to 0. This choice ensures that when the drop-out process is ignorable, then $q(u, L) = 0$ and, thus, $h\left\{u, \overline{L}(u)\right\}$ can be directly interpreted as giving the dependence of drop-out process on the recorded past. Here we have used the fact that the data are MAR if and only if $q(u, L) = q\left(u, \overline{L}(u)\right)$.

REMARK 2.10. Consider a study in which one wants to make inferences about time to death. However, we do not observe each subject's death time because of censoring by administrative end of follow-up or drop-out. In order to use the above methods to estimate the time to death distribution, we cannot treat death as a censoring event but rather must include it in the full data. Formally, we can write $L(u) = (D(u), D(u) V(u))$ where $D(u) = 1$ if the subject is alive, $D(u) = 0$ if the subject has died, and $D(u) V(u)$ denotes other characteristics recorded at u in living persons. Let T be time to death, i.e., $D(T) = 0$ and $D(T^-) = 1$. Let Q be time to drop-out (censoring). Theorem 2.3 assumes that censoring time C is always observed. However, in studies with death as the survival time, the censoring time Q is often not known (observed) for deaths. However, such studies can be easily accommodated in our set-up by setting the censoring time to infinity for deaths. That is, we set $C = Q$ if $Q < \min(\tau, T)$ and set $C = \infty$ otherwise. The observed data are now in the required form $\left(C, \overline{L}(C)\right)$. This renaming trick so that the censoring time C is always

observed is quite generally applicable (Robins, 1996; Gill, van der Laan, and Robins, 1998). A similar renaming trick will be necessary in Sec. 2.2 when we are interested in mortality as an endpoint. If we are interested in only a single cause of death, we let T be the time to death from that cause and count deaths from other causes as censoring events.

3. Identification of a subcomponent distribution in monotone missing data problems.

3.1. Single occasion. Suppose that in the scenario of Section 2.1, $L = (Y, V)$. However, suppose that interest lies in making inference about some component of the law of Y. That is, in our analysis Y is the sole outcome of interest despite the fact that the study also collects data on a secondary outcome V. Suppose that we are interested in performing a sensitivity analysis only over the magnitude of Y's influence on selection. The following generalization of Theorem 2.1 provides the mathematical justification for the discussion that follows.

THEOREM 3.1. *Given (i), a law F_O of $O = (\Delta, \Delta L)$ satisfying*

$$pr(\Delta = 1) \neq 0, \tag{3.1}$$

(ii) a continuous, monotone increasing, distribution function $\Phi(x)$, and (iii) a known function $q(Y)$, there exists joint laws $F_{\Delta,L}$ for (Δ, L) with marginal F_O for O satisfying

$$pr[\Delta = 1 \mid Y] = \Phi[h + q(Y)] \tag{3.2}$$

for some $h \in (-\infty, \infty]$. Specifically, h is the unique solution to

$$E_O[\Delta / \Phi(h + q(Y))] = 1 \tag{3.3}$$

and each law $F_{\Delta,L}$ has the same marginal F_Y for Y given by

$$F_Y(y) = E_O[\Delta I(Y \leq y) / \Phi\{h + q(Y)\}] \ . \tag{3.4}$$

Consider the semiparametric model **b** for the law of (Δ, L) defined by the following conditions: *(i)* the law of F_Y is unrestricted, *(ii)* the law of Δ given Y is given by (3.2) with h unknown, but $\Phi(x)$ and $q(Y)$ known and *(iii)* the law of V given (Y, Δ) is unrestricted. This model is a nonparametric model for the observed data O in which the constant h and the marginal distribution of Y are just identified. Note, that for a given law F_O and $\Phi(x)$, F_Y based on model **b** will generally differ from that based on model **a** of Sec. 2.1 unless the function $q(L) = q(Y, V)$ chosen in specifying model **a** equals the function $q(Y)$ chosen in specifying model **b**. From the point of view of model **a**, choosing $q(Y, V)$ to be a function of Y only is equivalent to specifying that Δ and V are conditionally independent given Y, an assumption that cannot be checked from the data since any choice

of $q(Y, V)$ is perfectly compatible with the distribution F_O of the observed data.

We now provide a generalization of Theorem 3.1 to the longitudinal monotone missing data setting of Section 2.2. Suppose that $L_k = (Y_k, V_k), k = 0, \ldots, K+1$, and we are interested in making inference about some component of the law F_Y of an outcome of interest $Y = (Y_0, \ldots, Y_{K+1})$. In particular, we want to conduct a sensitivity analysis over the influence of the magnitude of the unobserved parts of Y on the conditional probability of selection at each occasion $k+1$ given the observed past $\overline{L}_k \equiv (L_0, \ldots, L_k)$ and the current and future outcomes $\underline{Y}_{k+1} \equiv (Y_{k+1}, \ldots, Y_{K+1})$. Let $\lambda_k\left(\overline{L}_k, \underline{Y}_{k+1}\right)$ denote $pr\left(C = k \mid C \geq k, \overline{L}_k, \underline{Y}_{k+1}\right)$. On occasions, we write Λ_k for $\lambda_k\left(\overline{L}_k, \underline{Y}_{k+1}\right)$. Note that Λ_k is defined as a conditional probability given $\overline{L}_k$ and $\underline{Y}_{k+1}$, i.e., the observed past and the future outcomes of interest Y. In contrast, in Section 2.2, Λ_k was a conditional probability given the entire vector $(L_0, \ldots, L_{K+1})$. Theorem 3.1 has the following generalization formulated here directly in terms of a semiparametric model **b** to avoid redundancy. A proof follows along the lines of Lemma A.1 of Scharfstein, Rotnitzky, and Robins (1999).

THEOREM 3.2. *Consider the semiparametric model* **b** *for* (C, L) *characterized by the sole restrictions*

$$1 - \Lambda_k \equiv pr\left[C \neq k \mid C \geq k, \overline{L}_k, \underline{Y}_{k+1}\right] = \Phi\left[h_k\left(\overline{L}_k\right) + q_k\left(\overline{L}_k, \underline{Y}_{k+1}\right)\right] \tag{3.5}$$

$\Phi(\cdot)$ *a continuous, monotone increasing, distribution function,* $q_k(\cdot, \cdot)$ *a known function, and* $h_k(\cdot)$ *an unknown function taking values in* $(-\infty, \infty]$. *Suppose*

$$pr\left(C = k \mid C \geq k, \overline{L}_k\right) \neq 1 \ w.p.1 \ . \tag{3.6}$$

Then model **b** *is a non-parametric model for the law* F_O *of the observed data* $O = \left(C, \overline{L}_C\right)$. *Furthermore,* F_Y *and the* $h_k\left(\overline{L}_k\right)$ *are identified from data on* O. *Specifically,* $h_m(\cdot), m = K, \ldots, 0$, *are obtained as the unique solutions to the recursive set of equations*

$$E_O\left[\Delta / \left\{ \prod_{j=m+1}^{K} (1 - \Lambda_j)\, \Phi\left[h_m\left(\overline{L}_m\right) + q_m\left(\overline{L}_m, \underline{Y}_{m+1}\right)\right] \right\} \mid \overline{L}_m, C \geq m \right] = 1 \tag{3.7a}$$

for $m = K, \ldots, 0$, *where* $\prod_{j=a}^{b} V_j \equiv 1$ *if* $a > b$, *and*

$$F_Y(y) = E_O\left[\Delta I(Y \leq y) / \prod_{j=0}^{K} (1 - \Lambda_j)\right] . \tag{3.7b}$$

Furthermore,

$$F_{\underline{Y}_{m+1}}\left(\underline{y}_{m+1}\overline{L}_m, C \geq m\right) = E_O\left[\Delta I\left(\underline{Y}_{m+1} \leq \underline{y}_{m+1}\right) / \prod_{j=m}^{K}\left(1-\Lambda_j\right)|\overline{L}_m, C \geq m\right]. \tag{3.7c}$$

$$F_{\underline{Y}_{m+1}}\left(\underline{y}_{m+1} \mid \overline{L}_m, C > m\right) = E_O\left[\Delta I\left(\underline{Y}_{m+1} \leq \underline{y}_{m+1}\right) / \prod_{j=m+1}^{K}\left(1-\Lambda_j\right)|\overline{L}_m, C > m\right]. \tag{3.7d}$$

The key difference between model **a** of Section 2.2 and model **b** of this section is that model **b** imposes a restriction on the conditional probability of response at occasion $k+1$ given $\left(\overline{L}_k, \underline{Y}_{k+1}\right)$ while model **a** imposes a restriction on the conditional probability of response at $k+1$ given $\left(\overline{L}_k, \underline{L}_{k+1}\right)$. Inference about functionals of the law of F_Y under these two models will, quite generally, differ. To see this, notice that with $\Lambda_k \equiv pr\left(C = k \mid C \geq k, \overline{L}_k, \underline{Y}_{k+1}\right)$, as in this section, the density of the observed data $\left(C, \overline{L}_C\right)$ satisfies

$$f\left(C, \overline{L}_C\right) = \int f(Y) \prod_{m=0}^{C} f\left(V_m \mid Y, \overline{V}_m, C \geq m\right)\left\{\prod_{m=0}^{C-1}\left(1-\Lambda_m\right)\right\} \Lambda_C^{I(C \neq K+1)} d\mu\left(\underline{Y}_{C+1}\right) \tag{3.8}$$

while with $\Lambda_k \equiv pr\left(C = k \mid C \geq k, \overline{L}_k, \underline{L}_{k+1}\right)$, as in Sec. 2.2, the density of the observed data satisfies

$$f\left(C, \overline{L}_C\right) = \int f(Y) \prod_{m=0}^{C} f\left(V_m \mid Y, \overline{V}_m\right)\left\{\prod_{m=0}^{C-1}\left(1-\Lambda_m\right)\right\} \Lambda_C^{I(C \neq K+1)} d\mu\left(\underline{Y}_{C+1}\right) \tag{3.9}$$

where by convention we set $\underline{L}_{K+2} = 0$ and $\overline{V}_{-1} = 0$. Then, given an observed data law $f\left(C, \overline{L}_C\right)$, we conclude that $pr\left(C = k \mid C \geq k, \overline{L}_k, \underline{Y}_{k+1}\right) \neq pr\left(C = k \mid C \geq k, \overline{L}_k, \underline{L}_{k+1}\right)$ for some k (i.e., Λ_k defined in this section differs from Λ_k of Sec. 2.2) if and only if the unique density $f(Y)$ that solves the integral equation (3.8) under model **b** will not be the same as the unique density $f(Y)$ that solves the integral equation (3.9) under model **a**. The two versions of Λ_k will be equal if in model **a** we choose functions $q_k(L)$ that: *(i)* do not depend on $\underline{V}_{k+1} = (V_{k+1}, \ldots, V_{K+1})$, and *(ii)* are identically equal to the function $q_k\left(\overline{L}_k, \underline{Y}_{k+1}\right)$ that we choose in model **b**.

We now proceed to generalize those results to the case in which censoring is measured in continuous time. Henceforth, suppose that in

the scenario of Sec. 2.3, $L(u) = (Y(u), V(u))$. Define $\underline{Y}(u) \equiv \{Y(t); u \leq t \leq \tau\}$ and let $\lambda\left(u \mid \underline{Y}(u), \overline{L}(u^-)\right)$ be the conditional hazard of censoring at time u given $\underline{Y}(u)$ and $\overline{L}(u^-)$. On occasions, we write $\Lambda(u)$ for $\lambda\left(u \mid \underline{Y}(u), \overline{L}(u^-)\right)$.

We obtain an analogous result with censoring in continuous time.

THEOREM 3.3. *Consider the semiparametric model* $\boldsymbol{b}$ *for* (C, L) *characterized by the sole restriction on the conditional hazard of* C *given* $(\overline{L}(u^-), \underline{Y}(u) = \{Y(t); u \geq t \geq \tau\})$

$$\begin{aligned} \Lambda(u) &\equiv \lambda\left(u \mid \underline{Y}(u), \overline{L}(u^-)\right) \\ &= \exp\left[h\left(u, \overline{L}(u^-)\right) + q\left(u, \underline{Y}(u), \overline{L}(u^-)\right)\right] \end{aligned} \tag{3.10}$$

with $h\left(u, \overline{L}(u^-)\right)$ *unknown, and* $q\left(u, \underline{Y}(u), \overline{L}(u^-)\right)$ *known. Suppose* $\lambda\left(u \mid \overline{L}(u^-)\right)$ *is bounded w.p.*1. *Then model* $\boldsymbol{b}$ *is a non-parametric model for the law* F_O *of* $O = \left(C, \overline{L}(C)\right)$ *and* $h\left(u, \overline{L}(u)\right)$ *and* F_Y *are identified from data on* O. *Specifically,* $h\left(u, \overline{L}(u)\right)$ *is the unique function satisfying (2.14) with* $\Lambda(x)$ *redefined as in (3.10) and*

$$F_Y(y) = E_O\left[\Delta I(Y \leq y)/S\right] \tag{3.11}$$

where S *is as in (2.16) except with* $\Lambda(x)$ *as redefined in (3.10).*

Theorem 3.3 is proved in Lemma A.1 of Scharfstein, Rotnitzky, and Robins (1999) in the special case where the $\overline{L}(u)$ process jumps only at a finite number of fixed non-random times.

4. Estimation in monotone missing data problems.

4.1. Estimation — Simple example. For brevity, we will consider only the most difficult case, that of longitudinal monotone missing data in continuous time.

We consider estimation of functionals $\beta = \beta(F_L)$ of the distribution F_L from n i.i.d. copies of data $O = \left(C, \overline{L}(C)\right)$. We limit, for the moment, consideration to functionals β that admit unbiased estimating functions $U \equiv U(\beta) = u(L, \beta)$ for β in a semiparametric model in which F_L is unrestricted (non-parametric) but L is always fully observed. When F_L is unrestricted, $U(\beta)$ is unique up to multiplicative constants. For example, if $\beta = E[Y(10)]$, the mean at $t = 10$ of a variable $Y(t)$ that is a component of $L(t)$, then $U(\beta) = Y(10) - \beta$. The functionals β we are considering are those with positive semiparametric information bound when F_L is unrestricted and data on L is available.

If L was always observed, we would estimate β by $\widetilde{\beta}$ solving $\widetilde{E}_n[U(\beta)] = 0$ where $\widetilde{E}_n(H) = n^{-1}\sum_{i=1}^n H_i$ is the sample average over the n observations. For example, if $\beta = E[Y(10)]$, $\widetilde{\beta}$ is the sample average $\widetilde{E}_n[Y(10)]$.

We shall now consider estimation of β from data $O = \left(C, \overline{L}(C)\right)$ both in the semiparametric model $\mathbf{a}$ of Remark (2.8) and the more *restrictive* semiparametric model $\mathbf{a}_r$ in which we assume the unknown function $h\left(u, \overline{L}(u)\right)$ satisfies

$$\exp\left\{h\left(u,\overline{L}\left(u^{-}\right)\right)\right\}=\exp\left[h\left(u\right)+\nu\left(u,\overline{L}\left(u\right);\gamma_{0}\right)\right] \tag{4.1}$$

where $h(u)$ is an unknown function of time, $\nu(\cdot,\cdot,\cdot)$ is a known function, and γ_0 is the finite dimensional parameter to be estimated.

REMARK 4.1. In view of the discussion following Theorem 3.2, the results obtained in this section for estimation of model **a** apply equally to model **b**. Similarly, the results we obtained for estimation of $\mathbf{a_r}$ apply to model $\mathbf{b_r}$ where model $\mathbf{b_r}$ is model **b** with the additional restriction (4.1). This reflects the fact that for the purposes of semiparametric statistical inference, we can consider model **b** to be the special case of model **a** in which the known selection bias function $q(u, L)$ depends on L only through $\left(\underline{Y}(u),\overline{L}\left(u^{-}\right)\right)$.

As described previously, we need to consider the submodel $\mathbf{a}_r$ of our model **a** because of the curse of dimensionality. Specifically, in general, $\sqrt{n}-$ estimation of β under model **a** requires that we construct an estimate $\widehat{h}\left(u,\overline{L}\left(u^{-}\right)\right)$ that converges to $h\left(u,\overline{L}\left(u^{-}\right)\right)$ at rate $n^{\frac{1}{4}}$ or more, which is not practically possible (even with multivariate non-parametric smoothing techniques) in moderate size samples when $\overline{L}(u)$ is a high-dimensional stochastic process (Robins and Ritov, 1997).

To describe how we estimate β from data $O=\left(C,\overline{L}\left(C\right)\right)$ under semiparametric model **a** (when $\overline{L}(u)$ is not high-dimensional) or under model $\mathbf{a}_r$, we need for the moment to consider, as a pedagogic tool, a semiparametric model $\mathbf{a}_p$ (p for pedagogic) in which data *(i)* O is observed, *(ii)* F_L is unrestricted, and *(iii)* $\Lambda(u)\equiv\lambda(u\mid L)$ is completely known. Robins and Rotnitzky (1992) show that under model $\mathbf{a}_p$, the set of all unbiased estimating functions for β is

$$\mathcal{N}_1^{O,\perp}=\left\{N_1^{O,\perp}\left(\beta\right)\equiv\Delta U\left(\beta\right)/S+N_{car}\ ;\ N_{car}\in\mathcal{N}_{car}\right\} \tag{4.2}$$

where S is given by (2.16) and

$$\begin{aligned}\mathcal{N}_{car}=e\ \{&N_{car}\equiv N_{car}\left(a\right)\equiv\left(1-\Delta\right)A-\Delta E\left[\left(1-\Delta\right)A\mid L\right]/S\ ;\\ &A=a\left(O\right)\ \textit{has finite variance}\}\ .\end{aligned} \tag{4.3}$$

The reason for the notation we have chosen will become clear in the next subsection. $\mathcal{N}_{car}$ constitute exactly all functions of the observed data O with mean zero given the full data L. Since $E\left[\Delta U\left(\beta\right)/S\mid L\right]=U\left(\beta\right)$ and, at the true value β_0 of β, $E\left[U\left(\beta_0\right)\right]=0$, it is clear that $E\left[N_1^{O,\perp}\left(\beta_0\right)\right]=0$.

In semiparametric model **a** and $\mathbf{a}_r$, $\Lambda(u)$ and thus S are unknown, so the unbiased estimating function $N_1^{O,\perp}(\beta)$ cannot be computed. In this subsection we consider model **a** and assume $\tau_i=\tau$ for all i. To obtain estimates $\widehat{h}\left(u,\overline{L}\left(u\right)\right)$ and thus $\widehat{\Lambda}\left(u\right)\equiv\widehat{\lambda}\left(u\mid L\right)$ of (2.13), we first renumber the subjects $i, i=1,\ldots,n$, such that if $C_i<C_j$, then $i<j$. Next set $\exp\left\{\widehat{h}\left(u,\overline{L}\left(u\right)\right)\right\}=0$ unless $u=C_i<\tau_i$ for some $i=1,\ldots,n$. For

$C_i < \tau_i$, recursively define $\exp\left\{\widehat{h}\left(C_i, \overline{L}_i\left(C_i\right)\right)\right\}$ starting from the largest C_i as

$$\left[\sum_{j=1}^{n} \Delta_j I\left[\overline{L}_j\left(C_i\right) = \overline{L}_i\left(C_i\right)\right] \exp\left[q\left(C_i, L_j\right)\right] \Big/ \left\{\prod_{k=i+1}^{n}\left[1 - \widehat{\Lambda}_j\left(C_k\right)\right] I\left(C_k \neq \tau_k\right)\right\}\right]^{-1}. \tag{4.4}$$

Letting $\widehat{N}_1^{O,\perp}(\beta)$ be $N_1^{O,\perp}(\beta)$ with $\widehat{\Lambda}$ substituted for Λ, the estimator $\widehat{\beta}$ solving $0 = \sum_i \widehat{N}_{1i}^{O,\perp}(\beta)$ will be $\sqrt{n}$ -consistent for β in model **a** only if $\sum_{j=1}^{n} I\left(\overline{L}_j\left(C_i\right) = \overline{L}_i\left(C_i\right)\right) \to \infty$ as $n \to \infty$ with probability 1, which is an unreasonable asymptotics in high-dimensional problems due to the curse of dimensionality. In practice we would use model $\mathbf{a}_r$ whose estimation we will consider in Sec. 4.2.

However, for pedagogic purposes, in this subsection, we shall unrealistically continue to assume here that $\overline{L}(u)$ is discrete with only a moderate number of levels for each u so that the estimator $\widehat{\beta}$ is a regular asymptotically linear (RAL) estimator of β in model **a**. Our goal now is to derive a consistent estimator for the asymptotic variance of $n^{\frac{1}{2}}\left(\widehat{\beta} - \beta_0\right)$ so we can construct Wald confidence intervals for β. Recall that an estimator $\widetilde{\beta}$ 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. with mean zero and finite variance 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 i.i.d. random variables IF_i and the asymptotic variance of $n^{\frac{1}{2}}\left(\widetilde{\beta} - \beta_0\right)$ is $E\left[IF^{\otimes 2}\right]$. Thus, the goal is to find the influence function of the estimator $\widehat{\beta}$ solving $\sum_i \widehat{N}_{1i}^{O,\perp}(\beta) = 0$ and then estimate $E\left[IF^{\otimes 2}\right]$ by $\tilde{E}_n\left[\widehat{IF}^{\otimes 2}\right]$ where $\widehat{IF}^{\otimes 2}$ is a consistent estimator of IF. Since we have seen that the semiparametric model **a** is a non-parametric model for the observed data O, it follows from Bickel et al. (1993) that all RAL estimators of β will have the same influence function. Hence all the $\widehat{\beta}$ will have the identical influence function irrespective of the choice of $A = a(O)$.

In any semiparametric model, the influence function IF of any RAL estimator *(i)* lies in the orthogonal complement to the nuisance tangent space of the model and *(ii)* satisfies $E\left[IFS_\beta^T\right] = Id$ where Id is the identity matrix, T denotes matrix transposition, and S_β is the score for β evaluated at the true distribution generating the data (Bickel et al., 1993). As just mentioned, in model **a**, there is a unique random variable satisfying both *(i)* and *(ii)*.

Our goal is thus to determine the orthogonal complement to the nuisance tangent space in model **a** which we do in the course of developing the general theory given in the following subsection.

4.2. General theory of estimation. Let $\left(C, L \equiv \overline{L}(\tau)\right)$ denote the complete data. Suppose we only observe $O = \left(C, \overline{L}(C)\right)$. Furthermore, we assume that *(i)* L follows an arbitrary semiparametric model, F_L, indexed by a px1 parameter β and an infinite dimensional parameter θ, and *(ii)* C given L follows an arbitrary semiparametric model, $F_{C|L}$, indexed by a qx1 parameter γ and an infinite dimensional parameter η. If the model $F_{C|L}$ is completely non-parametric, then there will be no parameter γ. We assume that the parameters in model F_L are (locally) variation independent of those in the model for $F_{C|L}$. We let β_0, γ_0, θ_0, and η_0 denote the true values of β, γ, θ, and η, respectively. We are interested in estimating $\psi_0 = (\beta_0', \gamma_0')'$. We observe n independent identically distributed copies of O_i.

We shall assume that the probability of complete observations is bounded away from zero. This condition is somewhat stronger than we actually need to obtain the results below, as discussed in Rotnitzky and Robins (1997). Specifically, we assume

$$S \equiv pr\left[\Delta = 1 \mid L\right] > \sigma > 0 \quad w.p.1 \text{ for some constant } \sigma.$$

Let $\mathcal{N}_1 = \mathcal{N}(F_L)$ and $\mathcal{N}_2 = \mathcal{N}(F_{C|L})$ denote the (nuisance) tangent spaces for θ and η, respectively had we observed (C, L). Throughout, all spaces are sub-spaces of the Hilbert space of $q + p$-dimensional mean zero random vectors with the covariance inner product computed under the truth. Note that $\mathcal{N}_1$ and $\mathcal{N}_2$ are orthogonal. For the "observed data", there is an induced semiparametric model which we denote by *Obs*. In model *Obs*, the observed data nuisance tangent space is $\mathcal{N}^O = \overline{\mathcal{N}_1^O + \mathcal{N}_2^O}$, where $\mathcal{N}_1^O$ is the observed data nuisance tangent space for θ and $\mathcal{N}_2^O$ is the observed data nuisance tangent space for η. Specifically, $\mathcal{N}_j^O = \overline{R(g \circ \Pi_j)}$, where $R(\cdot)$ is the range of an operator, $g : \Omega^{(C,L)} \to \Omega^{(O)}$ is the conditional expectation operator $g(\cdot) = E[\cdot|O]$, $\Omega^{(C,L)}$ and $\Omega^{(O)}$ are the spaces of all $p + q$ dimensional random functions of (C, L) and O respectively, Π_j is the Hilbert space projection operator from $\Omega^{(C,L)}$ onto $\mathcal{N}_j$ and $\overline{\mathcal{S}}$ denotes the close linear span of the set $\mathcal{S}$ (Bickel et al., 1993). A space superscripted by $\perp$ denotes the orthogonal complement of that space. We are interested in finding $\mathcal{N}^{O,\perp}$ because, in sufficiently smooth models, the set of influence functions of all asymptotically linear (RAL) estimators of ψ_0 is the set $\left\{E\left[\mathcal{A}S_\psi'\right]^{-1}\mathcal{A}; \mathcal{A} \in \mathcal{N}^{O,\perp}\right\}$ where S_ψ is the score for ψ evaluated at the truth. Another motivation for our interest in this space is as follows. An element in $\mathcal{N}^{O,\perp}$ is a $(p + q)$ dimensional function of the observed data and of the true values of the parameters, ψ_0, θ_0, and η_0. Denote this function by $N^{O,\perp} \equiv N^{O,\perp}(\psi_0, \theta_0, \eta_0)$. Suppose we estimate ψ_0 by $\widehat{\psi}$ solving $\sum_i N_i^{O,\perp}\left(\psi, \widehat{\theta}(\psi), \widehat{\eta}(\psi)\right) = 0$ where $\widehat{\theta}(\psi_0)$ and $\widehat{\eta}(\psi_0)$ converge to θ_0 and η_0, respectively. Then Bickel et al. (1993), van der Vaart (1991), and Newey (1990) show that under suitable regu-

larity conditions $\widehat{\psi}$ is a RAL estimator with influence function $\rho^{-1}N^{O,\perp}$ where $\rho = E\left[N^{O,\perp}S'_{\psi}\right] = -\partial E\left[N^{O,\perp}(\psi,\theta_0,\eta_0)\right]/\partial\psi\,|_{\psi=\psi_0}$. But this is the same influence function as would have been obtained by solving the estimating equation $\sum_i N_i^{O,\perp}(\psi,\theta_0,\eta_0) = 0$ in which the infinite dimensional components (θ_0,η_0) are known rather than estimated. It is precisely the orthogonality of $N^{O,\perp}$ to $\mathcal{N}^O$ which obviates the need to adjust the asymptotic variance for estimation of the nuisance parameters (θ_0,η_0).

Taking orthogonal complements, we can express $\mathcal{N}^{O,\perp}$ as $\mathcal{N}_1^{O,\perp} \cap \mathcal{N}_2^{O,\perp}$. Let $a(L)$ and $b(O)$ be $p+q$ dimensional functions of L and O, respectively. $\mathcal{N}_1^{O,\perp}$ has the interpretation as the orthogonal complement to the nuisance tangent space $\mathcal{N}_1^O$ in the semiparametric model in which $\lambda[u \mid L]$ (i.e., the law of $C \mid L$) is known.

Rotnitzky and Robins (1997) showed how to compute $\mathcal{N}_1^{O,\perp}$. Specifically,

$$(4.5)\quad \mathcal{N}_1^{O,\perp} = \left\{N_1^{O,\perp} = \Delta m(L)/S + N_{car} : m(L) \in \mathcal{N}_1^{\perp} \text{ and } N_{car} \in \mathcal{N}_{car}\right\}.$$

REMARK 4.2. Recall that N_{car} was defined in Eq. (4.3). $\mathcal{N}_{car}$ is exactly the nuisance tangent space (i.e., mean square closure of nuisance scores) corresponding to $\lambda(u \mid L)$ when $\lambda(u \mid L)$ is unrestricted except for the fact that the data are CAR, i.e., $q(u,L) = q\left(u,\overline{L}(u^-)\right)$.

By the relationship between range and null spaces, $\mathcal{N}_2^{O,\perp} = Null\left(\Pi_2^T \circ g^T\right)$, where $Null(\cdot)$ is the null space of an operator, and superscript T denotes the adjoint of an operator. As projection operators $\Pi_j^T = \Pi_j, j = 1,2$, and $g^T : \Omega^O \to \Omega^{(C,L)}$ is the identity operator. So,

$$(4.6)\qquad \mathcal{N}_2^{O,\perp} = \{b(O) : \Pi_2[b(O)|\mathcal{N}_2] = 0\} = \{b(O) : b(O) \in \mathcal{N}_2^{\perp}\}.$$

Model a: Our next goal is to determine the orthogonal complement to the nuisance tangent space in semiparametric model **a** of Remark 2.8, except we no longer require that F_L be unrestricted. Rather, we allow F_L to follow a semiparametric model indexed by finite dimensional parameter β and infinite dimensional parameter θ. Under model **a**, $F_{C|L}$ is given by (2.13) with $h\left(u,\overline{L}(u)\right)$ unrestricted and $q(u,L)$ known. In model **a**, there is no parameter γ so that $\psi = \beta$.

If we had observed (C,L), then the nuisance scores corresponding to parametric submodels for the unknown $h\left(u,\overline{L}(u)\right)$ is the set

$$(4.7)\qquad \mathcal{N}_2 = \left\{N_2 = N_2(a) = \int_0^{\tau} d\mathcal{M}(u)\, a\left(u,\overline{L}(u)\right)\right\}$$

where $a\left(u,\overline{L}(u)\right)$ is an arbitrary function of dimension equal to that of β, $\mathcal{M}(u) = I[C \le u, C < \tau] - \int_0^u \lambda(x \mid L)\, I(C \ge x)\, dx$ is the martingale for censoring conditional on L. Note if $C = \tau$, $I(C \le u, C < \tau) = 0$ since we

do not regard a subject successfully reaching end of follow-up as having been censored.

REMARK 4.3. Note that the N_2 are not necessarily functions of the observed data $O = \left(C, \overline{L}(C)\right)$. However, if the selection function $q(u, L) = q\left(u, \overline{L}(u)\right)$ w.p.1 so that the data are CAR, then $\mathcal{N}_2 = \mathcal{N}_{car}$ and the N_2 are functions of O.

The orthogonal complement to $\mathcal{N}_2$ of (4.7) is

$$\mathcal{N}_2^{\perp} = \left\{ N_2^{\perp} = \int_0^{\tau} d\mathcal{M}(u)\, b(u, L) + m(L)\,;\, E[m(L)] = 0 \;\; \textit{and} \right.$$
$$\left. E\left[b(u, L)\, S(u) \exp[q(u, L)] \mid \overline{L}(u)\right] = 0 \;\; \textit{for } u \in [0, \tau] \right\} \tag{4.8}$$

by Ritov and Wellner (1988) and Robins and Rotnitzky (1992). Here

$$S(u) \equiv \exp\left[-\int_0^u \lambda(s \mid L)\, ds\right] . \tag{4.9}$$

Given $\mathcal{N}_2^{\perp}$, we obtain $\mathcal{N}_2^{O,\perp}$ from (4.6).

For concreteness, we shall consider the semiparametric model for F_L in which the conditional mean of $Y \equiv Y(10)$ given a vector $X \in L(0)$ follows the parametric regression model

$$E[Y \mid X] = g(X, \beta_0) \tag{4.10}$$

where $g(X, \beta)$ is a known function. Note the non-parametric model for the mean of Y discussed in the previous subsection is a special case of model (4.10) in which X is constant with probability 1 and $g(X, \beta_0) = \beta_0$. That is,

$$E(Y) = \beta_0. \tag{4.11}$$

Let $\varepsilon \equiv Y - g(X, \beta_0)$. In the semiparametric model for F_L given by the sole restriction (4.10), Robins et al. (1994) showed that the orthogonal complement $\mathcal{N}_1^{\perp}$ to the nuisance tangent space $\mathcal{N}_1$ is

$$\mathcal{N}_1^{\perp} = \{U = d(X)\,\varepsilon\,;\; d(X) \;\; \textit{arbitrary}\} . \tag{4.12}$$

In the non-parametric model (4.11),

$$\mathcal{N}_1^{\perp} = \{U = c(Y - \beta_0)\,;\; c \textit{ is an arbitrary constant}\} . \tag{4.13}$$

For each $m(L) \in \mathcal{N}_1^{\perp}$, let $a_m\left(u, \overline{L}(u)\right)$ be the unique solution on $0 \leq u < \tau$ to the Volterra integral equation

$$a\left(u, \overline{L}(u)\right) = J_m(u) - \int_0^u ds\; a\left(s, \overline{L}(s)\right) \kappa\left(s, u, \overline{L}(u)\right) \tag{4.14}$$

where

$$J_m(u) \equiv E\left[m(L)\exp\{q(u,L)\} \mid \overline{L}(u)\right] \Big/ E\left[S(u)\exp\{q(u,L)\} \mid \overline{L}(u)\right] , \tag{4.15}$$

with $S(u)$ as defined in (4.9) and

$$\kappa\left(s,u,\overline{L}(u)\right) \equiv E\left[\Lambda(s)S(s)e^{q(u,L)} \mid \overline{L}(u)\right] \Big/ E\left[S(u)\exp\{q(u,L)\} \mid \overline{L}(u)\right] . \tag{4.16}$$

THEOREM 4.1. *In the semiparametric model* **a** *with* $\mathcal{N}_1^{\perp} = \{m(L)\}$

$$\mathcal{N}^{O,\perp} = \left\{N^{O,\perp} = N^{O,\perp}(m) = \Delta m(L)/S + N_{car}(a_m);\ m(L) \in \mathcal{N}_1^{\perp}\right\} \tag{4.17}$$

Proof. To prove the theorem, we use the following lemma, which we prove following the proof of Theorem 4.1.

LEMMA 4.1. *For any* $m(L)$ *and any* $A = a(O) \equiv a\left(C, \overline{L}(C)\right)$,

$$\Delta m(L)/S + (1-\Delta)A - \Delta E[(1-\Delta)A \mid L]/S = m(L) + \int d\mathcal{M}(u)\, b(u,L) \tag{4.18}$$

where

$$b(u,L) \equiv a\left(u,\overline{L}(u)\right) + E\left[a\left(C,\overline{L}(C)\right)I(C<u) \mid L\right]/S(u) - m(L)/S(u) . \tag{4.19}$$

Proof of Theorem 4.1. Theorem 4.1 follows by noting that it follows from the characterization of $\mathcal{N}_2^{\perp}$ in (4.8) that the LHS of (4.18) is in $\mathcal{N}_2^{\perp}$ if and only if for $b(u,L)$ given by (4.19), $E\left[b(u,L)S(u)e^{q(u,L)} \mid \overline{L}(u)\right] = 0$ which implies that

$$\begin{aligned} a(u,\overline{L}(u))E\left[S(u)\exp q(u,L) \mid \overline{L}(u)\right] - E\left[m(L)\exp\{q(u,L)\} \mid \overline{L}(u)\right] \\ + E\left[E\{a(C,\overline{L}(C))I(C<u) \mid L\}\exp\left\{q(u,L) \mid \overline{L}(u)\right\}\right] = 0 . \end{aligned} \tag{4.20}$$

The last term on the LHS of (4.20) is $\int_0^u ds\, a\left(s,\overline{L}(s)\right) E\left[\Lambda(s)S(s)\exp\{q(u,L)\} \mid \overline{L}(u)\right]$, which completes the proof.

Proof of Lemma 4.1.

$$\Delta m(L)/S = m(L) - \int_0^{\tau} m(L)\, d\mathcal{M}(u)/S(u) \tag{4.21}$$

by Eq. (3.10d) of Robins and Rotnitzky (1992). Furthermore,

$$(4.22)\qquad \int_0^\tau \{dN^*(u) - \Delta S(u) S^{-1}\lambda(u \mid L) I(C \geq u)\, du\} a\left(u, \overline{L}(u)\right) = (1-\Delta) A - \Delta E[(1-\Delta) A \mid L]/S$$

where $N^*(u) = I[C \leq u, C < \tau]$ since subjects with $\Delta = 0$ contribute $(1-\Delta) A$ to the LHS of (4.22), and the LHS of (4.22) has mean zero given L, since $E[\Delta S(u)/S \mid L, I(C \geq u)] = I(C \geq u)$ and $E[dN^*(u) \mid L, I(C \geq u)] = \lambda(u \mid L) I(C \geq u)\, du$. However, by (3.10c) of Robins and Rotnitzky (1992), $\{\Delta S(u)/S\}\lambda(u \mid L) I(C \geq u) = \lambda(u \mid L) I(C \geq u) - S(u)\lambda(u \mid L)\int_u^\tau d\mathcal{M}(x)/S(x)$. So the LHS of (4.22) is $\int_0^\tau d\mathcal{M}(u)\, a(u, \overline{L}(u)) + \int_0^\tau duS(u)\lambda(u \mid L)\, a\left(u, \overline{L}(u)\right)\int_u^\tau d\mathcal{M}(x)/S(x)$.

However, by Fubini's theorem, $\int_0^\tau duS(u)\Lambda(u)\, a\left(u, \overline{L}(u)\right)\int_u^\tau d\mathcal{M}(x)/S(x) = \int_0^\tau d\mathcal{M}(x)\{S(x)\}^{-1}\int_0^x du\left(S(u)\Lambda(u)\, a\left(u, \overline{L}(u)\right)\right) = \int_0^\tau d\mathcal{M}(x) E\left[I(C < x)\, a\left[C, \overline{L}(C)\right] \mid L\right]/S(x)$ which proves the lemma. □

REMARK 4.4. In model **a**, with $\beta_0 = E[Y(10)]$ and F_L unrestricted, $\mathcal{N}_1^\perp$ is given by Eq. (4.13). It then follows from Theorem 4.1 that $\mathcal{N}^{O,\perp}$ is comprised of multiples of a single random variable. Hence all RAL estimators of β_0 must have the same influence function.

In model **a** in which F_L follows the semiparametric model characterized by (4.10), there is more than one influence function, and the question arises as to which influence function is optimal. Now the influence function $IF(m)$ associated with an element $N^{O,\perp}(m)$ in $\mathcal{N}^{O,\perp}$ is

$$(4.23)\qquad IF(m) = E\left[N^{O,\perp}(m)\, S_\beta^T\right]^{-1} N^{O,\perp}(m)$$

where S_β is the score for β evaluated at the truth. However, Robins, Rotnitzky, and Zhao (1994) show that

$$(4.24)\qquad E\left[N^{O,\perp}(m)\, S_\beta^T\right]^{-1} = E\left[m(L)\, S_{eff}^{F,T}\right]^{-1}$$

where $S_{eff}^F = \Pi\left[S_\beta^F \mid \mathcal{N}_1^\perp\right]$ is the efficient score for $\beta_0 = \psi_0$ were the full data (C, L) always observed. Here, and throughout: $\Pi(A \mid \mathcal{A})$ denotes the Hilbert space projection of the random variable A on the space $\mathcal{A}$, and S_β^F is the score for β when data on (C, L) are available.

The efficient influence function, $EIF \equiv IF(m_{eff})$ has minimum variance among all members of the set of influence functions $\{IF(m)\}$. $Var[IF(m_{eff})]$ is the semiparametric variance bound for model **a**. The following lemma is an immediate consequence of (4.23), (4.24), and Theorem 5.3 in Newey and McFadden (1993).

LEMMA 4.2. *$m_{eff}(L)$ is the unique member of $\mathcal{N}_1^\perp$ that satisfies*

$$(4.25)\qquad E\left[m(L)\, S_{eff}^{FT}\right] = E\left[N^{O,\perp}(m)\, N^{O,\perp}(m_{eff})^T\right]$$

for all $m(L) \in \mathcal{N}_1^\perp$.

COROLLARY 4.1. *$m_{eff}(L)$ is the unique member of $\mathcal{N}_1^{\perp}$ satisfying*

$$S_{eff}^F = \Pi\left[\mathbf{O}^{\dagger}\mathbf{O}\left[m_{eff}(L)\right] \mid \mathcal{N}_1^{\perp}\right]$$

where the operator $\mathbf{O}: \Omega^{(L)} \to \Omega^{(O)}$ maps $m(L)$ into $\mathcal{N}^{O,\perp}(m)$ and $\mathbf{O}^{\dagger}: \Omega^{(O)} \to \Omega^{(L)}$ is the adjoint of $\mathbf{O}$.

Consider model **a** with F_L restricted by (4.10). Chamberlain (1987) shows that $S_{eff}^F = \{\partial g(X,\beta_0)/\partial\beta\}\, var(\varepsilon \mid X)^{-1}\varepsilon$. Further, by (4.12), $m(L) \in \mathcal{N}_1^{\perp} \Leftrightarrow m(L) = d(X)\varepsilon$. Thus the left hand side of (4.25) becomes $E\left[d(X)\{\partial g(X,\beta_0)/\partial\beta\}^T\right]$. The RHS of (4.25) becomes $E\left[d(X)\, d_{eff}(X)\, \kappa^*(X)\right]$ where $\kappa^*(X) = E\left[\{\Delta\varepsilon/S + N_{car}(a_{eff})\}^2 \mid X\right]$, $m_{eff}(X) \equiv d_{eff}(X)\varepsilon$, $a_{eff}\left(u, \overline{L}(u)\right)$ is the solution $a_m\left(u, \overline{L}(u)\right)$ to (4.14) with $m(L) \equiv \varepsilon \equiv Y - g(x,\beta_0)$. [We have used the fact that, for $m(L) = d(X)\varepsilon$, $a_m\left(u, \overline{L}(u)\right) = d(X)\, a_{eff}\left(u, \overline{L}(u)\right)$.] The LHS and RHS of (4.25) must be equal for all $d(X)$. This implies that

$$d_{eff}(X) = \{\partial g(X,\beta_0)/\partial\beta\}/\kappa^*(X)\ . \tag{4.26}$$

4.3. Further details of estimation in model a. We continue to consider semiparametric model **a** with F_L following the semiparametric model (4.10). In practice we will estimate β_0 by $\widehat{\beta}(d)$ solving $0 = \sum_i \widehat{N}_{1i}^{O,\perp}(\beta, d) = 0$ where

$$\widehat{N}_1^{O,\perp}(\beta, d) = \Delta d(X)\,\varepsilon(\beta)/\widehat{S} + \widehat{N}_{car}\ , \tag{4.27}$$

$\widehat{N}_{car} \equiv \widehat{N}_{car}(a) = (1-\Delta)A - \Delta E\left[(1-\Delta)A \mid L\right]/\widehat{S}$, $\widehat{\Lambda}(u)$, $\varepsilon(\beta) = Y - g(X,\beta)$, and $\widehat{E}[\cdot \mid L]$ and $\widehat{S}$ are determined by the estimates based on (4.4). Under mild regularity conditions for each choice of $d(X)$, $\widehat{\beta}(d)$ will be a RAL estimator and the asymptotic variance of $n^{\frac{1}{2}}\left(\widehat{\beta}(d) - \beta_0\right)$ can be estimated using the bootstrap (Gill, 1989). However, the analytic sandwich estimator of the asymptotic variance

$$\widehat{I}^{-1}(d)\left[n^{-1}\sum_i \widehat{N}_{1i}^{O,\perp}(\beta, d)^{\otimes 2}\right]\widehat{I}(d)^{-1T} \tag{4.28}$$

evaluated at $\beta = \widehat{\beta}(d)$ with $\widehat{I}(d) = n^{-1}\sum_i \partial\widehat{N}_{1i}^{O,\perp}(\beta, d)/\partial\beta$ will be inconsistent for the asymptotic variance of $n^{\frac{1}{2}}\left(\widehat{\beta}(d) - \beta_0\right)$ unless A is equal to $A_m = a_m\left(C, \overline{L}(C)\right)$ solving (4.14) with $m(L) = d(X)\varepsilon$ (so that $N_1^{O,\perp}(\beta_0, d)$ is in $\mathcal{N}^{O,\perp}$ as well). However, the above analytic estimator can be used if we replace A by a consistent estimator $\widehat{A}_m$ of A_m obtained by solving the Volterra integral equation (4.14) with *(i)* $\widehat{\Lambda}(s)$ and $\widehat{S}(u)$ based on (4.4) replacing $\Lambda(s)$ and $S(s)$ and *(ii)* for any random H, $E\left[H \mid \overline{L}(u)\right]$ is consistently estimated, by $\left\{\sum_i I\left[\overline{L}_i(u) = \overline{L}(u)\right]\Delta_i H_i \widehat{S}_i(u)/\widehat{S}_i\right\}$ $/\left\{\sum_i I\left[\overline{L}_i(u) = \overline{L}(u)\right]\Delta_i \widehat{S}_i(u)/\widehat{S}_i\right\}$. Replacement of A_m by $\widehat{A}_m$ does not

change the asymptotic distribution of the estimator. Note that the Volterra integral equation (4.14) becomes a finite dimensional matrix equation in its estimated form. (Since we are continuing to assume that $\overline{L}(u)$ has only a few levels, the expectations from the Volterra integral equation can be estimated by sample averages.) Finally, a semiparametric efficient estimator can be obtained by estimating $d_{eff}(X)$ based on a consistent preliminary estimators of β_0.

4.4. Estimation in model $\mathbf{a}_r$. Model $\mathbf{a}_r$ differs from model $\mathbf{a}$ by imposing (4.1). As in Sec. 2, we allow F_L to follow a semiparametric model indexed by finite dimensional parameter β and infinite dimensional parameter θ. Thus in model $\mathbf{a}_r$, $\psi = (\beta', \gamma')'$ is the parametric component. The infinite dimensional component η associated with $F_{C|L}$ indexes functions $h(u)$ of u. If $h(u)$ were known, we would consider estimating equations based on $N_1^{O,\perp} \equiv N_1^{O,\perp}(\psi_0)$ of (4.5) indexed by $p+q$ dimensional functions $m(L) \in \mathcal{N}_1^{\perp}$ and $a(O) \in \mathcal{N}_{car}$. That is, we would solve $0 = \sum_i N_{1i}^{O,\perp}(\psi)$. Since $h(u)$ is unknown, we instead solve $0 = \sum_i \widehat{N}_{1i}^{O,\perp}(\psi)$ in which an estimator $\widehat{h}(u,\gamma)$ of $h(u)$ depending on γ has been substituted in $N_{1i}^{O,\perp}(\psi)$. For example, if F_L follows the model characterized by (4.10), $\widehat{N}_{1i}^{O,\perp}(\psi) = \Delta d(X)\varepsilon(\beta)/\widehat{S}(\gamma) - \widehat{N}_{car}(a,\gamma)$ where $\widehat{N}_{car}(a,\gamma)$ and $\widehat{S}(\gamma)$ are $N_{car}(a)$ and S with $\widehat{h}(u,\gamma)$ substituted for $h(u)$.

Specifically, let $Z(u,\gamma) = \exp\left[\nu\left(u, \overline{L}(u); \gamma\right) + q(u,L)\right]$ and again renumber subjects so that if $C_i < C_j$ that $i < j$. Then we recursively define estimators $\widehat{h}(u,\gamma)$ and thus $\widehat{\Lambda}(u;\gamma)$ as follows. Set $\exp\left\{\widehat{h}(u;\gamma)\right\} = 0$ unless $u = C_i < \tau_i$ for some i. For $C_i < \tau_i$, recursively define $\exp\left\{\widehat{h}(C_i;\gamma)\right\}$ starting from the largest C_i as $\left[\sum_{j=1}^{n} \Delta_j Z_j(C_i;\gamma) \Big/ \prod_{k=i+1}^{n} [1 - \widehat{\Lambda}_j(C_k;\gamma)] I(C_k \neq \tau_k)\right]^{-1}$.

Our goal is now to determine the orthogonal complement to the nuisance tangent space in model $\mathbf{a}_r$ which includes all influence functions IF for ψ_0. In model $\mathbf{a}_r$,

$$\mathcal{N}_2 = \left\{ N_2 = N_2(a) = \int_0^r dM(u)\, a(u) \right\} . \tag{4.29}$$

Define $Z(u) = Z(u, \gamma_0)$. It follows by the results of Ritov and Wellner (1988) that

$$\mathcal{N}_2^{\perp} = \left\{ N_2^{\perp} = \int_0^{\tau} d\mathcal{M}(u)\, b(u,L) + m(L) \;;\; E[m(L)] = 0 \right. \\ \left. \text{and } E[b(u,L)\, S(u)\, Z(u)] = 0 \right\} . \tag{4.30}$$

Furthermore, $N_2^{O,\perp}$ is still defined in terms of $\mathcal{N}_2^{\perp}$ by (4.6), $\mathcal{N}_1^{O,\perp}$ is still given by (4.9) with $\mathcal{N}_1^{\perp}$ as in model $\mathbf{a}$ and $\mathcal{N}^{O,\perp}$ still $\mathcal{N}_1^{O,\perp} \bigcap \mathcal{N}_2^{O,\perp}$.

We now prove a lemma characterizing $\mathcal{N}^{O,\perp}$ in greater detail.

Given any function $m(L)$ and any function $a^*\left(u, \overline{L}(u)\right)$, let $a_{m,a^*}(u, \overline{L}(u))$ be the unique solution $a\left(u, \overline{L}(u)\right)$ to the "Volterra-like" recursive integral equation

$$
\begin{aligned}
a\left(u, \overline{L}(u)\right) &= a^*\left(u, \overline{L}(u)\right) \\
&- \left\{E\left[S(u) Z(u)\right]\right\}^{-1} \left\{ E\left[a^*\left(u, \overline{L}(u)\right) S(u) Z(u)\right] \right. \\
&\left. - E\left[m(L) Z(u)\right] + E\left[\left\{ \int_0^u a\left(s, \overline{L}(s)\right) \Lambda(s) S(s)\, ds \right\} Z(u) \right] \right\}.
\end{aligned} \tag{4.31}
$$

LEMMA 4.3. *In model* $\mathbf{a}_r$, $\mathcal{N}^{O,\perp} = \{N^{O,\perp} = N^{O,\perp}(m, a^*) = \Delta m(L) / S + N_{car}(a_{m,a^*})$; $a^*\left(u, \overline{L}(u)\right)$ *arbitrary,* $m(L) \in \mathcal{N}_1^{\perp}\}$.

Proof. We first follow the proof of Theorem (4.1) to obtain that the LHS of (4.18) is in $\mathcal{N}_2^{\perp}$ as given by (4.30) if and only if $b(u, L)$ given by (4.19) satisfies $E\left[b(u, L) S(u) Z(u)\right] = 0$. But this is true if and only if

$$
\begin{aligned}
&E\left[a\left(u, \overline{L}(u)\right) S(u) Z(u)\right] = \\
&\qquad E\left[m(L) Z(u)\right] - E\left[\left\{ \int_0^u a\left(s, \overline{L}(s)\right) \Lambda(s) S(s)\, ds \right\} Z(u)\right].
\end{aligned} \tag{4.32}
$$

Now $a(u, \overline{L}(u))$ solving (4.31) clearly satisfies (4.32). Conversely, if $a(u, \overline{L}(u))$ satisfies (4.32), then if we define $a^{\dagger}\left(u, \overline{L}(u)\right) = a\left(u, \overline{L}(u)\right) - \left\{E[m(L)Z(u)] - E[\{\int_0^u a(s, \overline{L}(s))\Lambda(s)S(s)ds\}Z(u)]\right\}/E\left(S(u)Z(u)\right)$, then $E\left[a^{\dagger}\left(u, \overline{L}(u)\right) S(u) Z(u)\right] = 0$. The Hilbert space projection of any function $a^*\left(u, \overline{L}(u)\right)$ on the space of functions $a^{\dagger}\left(u, \overline{L}(u)\right)$ satisfying the last equality is given by $a^*(u, \overline{L}(u)) - \{E[S(u)Z(u)]\}^{-1}E[a^*(u, \overline{L}(u))S(u)Z(u)]$ which proves the theorem. □

We now explain how to use Lemma 4.3 to construct asymptotic variance estimators. For concreteness, we continue to consider the model for F_L characterized by the sole restriction (4.10). Then the asymptotic variance of $n^{\frac{1}{2}}\left(\widehat{\psi}(d) - \psi_0\right)$ solving $0 = \sum_i \widehat{N}_{1i}^{O,\perp}(\psi, d)$ based on $d(X)\varepsilon \in N_1^{\perp}$ can be consistently estimated by (4.28) [with ψ replacing β] only if the term $N_{car}(a)$ in $N_1^{O,\perp}(\psi, d)$ has $a = a_{m,a^*}$ for $m(L) = d(X)\varepsilon$ and some $a^*\left(u, \overline{L}(u)\right)$ [so that $N_1^{O,\perp}(\psi_0, d)$ is also $N^{O,\perp}$]. However, the analytic variance estimator (4.28) can be used if we use $N_{car}(\widehat{a}_{m,a})$ where $\widehat{a}_{m,a^*}$ is a consistent estimator of a_{m,a^*} which we obtain by *(i)* calculating an estimator $\widetilde{\psi}$ of ψ_0 based on any initial $\widehat{N}_1^{O,\perp}(\psi, d)$ and then *(ii)* solving (4.31) with $\widehat{\Lambda}(s; \widetilde{\gamma})$ and $S(s; \widetilde{\gamma})$ replacing $\Lambda(s)$ and $S(s)$ and with inverse-probability-of-remaining-uncensored-weighted sample averages replacing expectations.

REMARK 4.5. We have not as yet characterized the efficient score in model $\mathbf{a}_r$. However, if as recommended in Remark 2.7, γ_0 is a high-

dimensional parameter and β_0 is the parameter of interest, then the semiparametric efficiency bound for β_0 in model $\mathbf{a}_r$ and model $\mathbf{a}$ should be similar. Thus we suggest using an estimator for β_0 that would be efficient in model $\mathbf{a}$. We can construct such an estimator, because we can know the form of the efficient estimator in model $\mathbf{a}$.

To clarify this proposal, suppose our model for F_L is still characterized by restriction (4.10). As with model $\mathbf{a}$, it is necessary to estimate $d_{eff}(X)$ of Eq. (4.26) and $a_{eff}\left(u, \overline{L}(u)\right)$ as defined in the paragraph preceding Eq. (4.26) from the data. This requires we solve an estimated version of the Volterra integral Equation (4.14). However, (4.14) requires that we compute conditional expectations given $\overline{L}(u)$ and X, both of which may now be high-dimensional and continuous. Thus, we suggest first one specify a fully parametric model for the joint distribution of (C, L), estimate the model by maximum likelihood from the observed data and construct estimators $\widehat{d}_{eff}(X)$ and $\widehat{a}_{eff}\left(u, \overline{L}(u)\right)$ based on the estimated law of (C, L). Then let $\widehat{a}_{m,a^*}$ be the estimated version of a_{m,a^*} calculated as described in the previous paragraph with $m(L) = \widehat{d}_{eff}(X)\,\varepsilon\left(\widetilde{\beta}\right)$, $a^*\left(u, \overline{L}(u)\right) = \widehat{d}_{eff}(X)\,\widehat{a}_{eff}\left(u, \overline{L}(u)\right)$ where $\widetilde{\beta}$ is a preliminary $n^{\frac{1}{2}}$ - consistent estimator of β_0. Finally, obtain $\widehat{\psi}(\widehat{d}_{eff}) = \left(\widehat{\beta}(\widehat{d}_{eff})^T, \widehat{\gamma}(\widehat{d}_{eff})^T\right)^T$ solving $0 = \sum_i \widehat{N}_{1i}^{O,\perp}\left(\psi, \widehat{d}\right)$ with the first p components of the $p+q$ - dimensional estimating function $N_{1i}^{O,\perp}\left(\psi, \widehat{d}\right)$ being $\Delta\widehat{d}_{eff}(X)\,\varepsilon(\beta)/\widehat{S}(\gamma) - \widehat{N}_{car}(\widehat{a}_{m,a^*}, \gamma)$ with $\widehat{a}_{m,a^*}$ as just defined. If the parametric model for (C, L) is correctly specified, then $\widehat{a}_{m,a^*}$ is consistent for $d_{eff}(X)\,a_{eff}(u, \overline{L}(u))$ and $\widehat{d}_{eff}(X)$ is consistent for $d_{eff}(X)$ [defined by Eq. (4.26)]. Furthermore, the solutions $\widehat{\beta}(\widehat{d}_{eff})$ and $\widehat{\gamma}(\widehat{d}_{eff})$ will be asymptotically independent and the asymptotic variance of $\widehat{\beta}(\widehat{d}_{eff})$ will attain the semiparametric variance bound for model $\mathbf{a}$ and thus, as argued above, be nearly efficient in model $\mathbf{a}_r$.

Even if the parametric model for (C, L) is misspecified, $\widehat{\psi}(\widehat{d}_{eff})$ is still a RAL estimator under model $\mathbf{a}_r$ with asymptotic variance that is consistently estimated by the analytic estimator (4.28) with β replaced by ψ. However, the asymptotic variance of $\widehat{\beta}(\widehat{d}_{eff})$ will no longer equal that of the efficient estimator under model $\mathbf{a}$, although one would expect that the difference would not be large provided the parametric model for (C, L) is richly parameterized.

5. Selection odds models and the selection bias G-computation algorithm formula.

5.1. Selection bias g-computation algorithm formula. In this subsection, we derive the non-ignorable selection bias g-computation algorithm formula. We return to the setting where drop-out occurs only at fixed times $k, k = 0, 1, \ldots, K$ and $L = \left(\overline{L}_{K+1}\right) = (L_0, \ldots, L_{K+1})$.

Define

$$
\begin{aligned}
&B_k\left(\underline{y}_{k+1}\right) = b_k\left(\underline{y}_{k+1}, \overline{L}_k\right) = f_{\underline{Y}_{k+1}}\left(\underline{y}_{k+1} \mid \overline{L}_k, C > k\right), \\
&B_k^*\left(\underline{y}_{k+1}\right) = b_k^*\left(\underline{y}_{k+1}, \overline{L}_k\right) = f_{\underline{Y}_{k+1}}\left(\underline{y}_{k+1} \mid \overline{L}_k, C \geq k\right) \text{ and} \\
&\Lambda_k^* = pr\left[C = k \mid C \geq k, \overline{L}_k\right].
\end{aligned} \tag{5.1}
$$

Consider now the selection bias g-computation algorithm formula identity

$$
\begin{aligned}
f_{\underline{Y}_{k+1}}\left(\underline{y}_{k+1} \mid \overline{\ell}_k, C > k\right) &\equiv b_k\left(\underline{y}_{k+1}, \overline{\ell}_k\right) \\
&= \int \cdots \iint \prod_{m=k+1}^{K+1} f\left[y_m \mid \overline{\ell}_{m-1}, C \geq m\right] dF\left[v_m \mid y_m, \right. \\
&\qquad \left. \overline{\ell}_{m-1}, C \geq m\right] \left\{ \prod_{m=k+1}^{K} j\left(\overline{\ell}_m, \underline{y}_{m+1}\right) \right\}
\end{aligned} \tag{5.2}
$$

with

$$
\begin{aligned}
j\left(\overline{L}_m, \underline{Y}_{m+1}\right) &\equiv b_m^*\left(\underline{Y}_{m+1}, \overline{L}_m\right) / b_m\left(\underline{Y}_{m+1}, \overline{L}_m\right) \\
&= \left(1 - \Lambda_k^*\right) / \left(1 - \Lambda_k\right),
\end{aligned} \tag{5.3}
$$

where the last identity is by Bayes' theorem and Λ_k is as in Theorem 3.2. It follows from (5.2) that to make $b_k\left(\underline{y}_{k+1}, \overline{\ell}_k\right)$ identifiable from data on O, we need to make the $j\left(\overline{\ell}_k, \underline{y}_{k+1}\right)$ or, equivalently, the Λ_k identifiable which, by Theorem 3.2, can be accomplished by imposing the semiparametric model b for (C, L) characterized by the restrictions (3.5), (3.6). Indeed, Eq. (5.2) is essentially Eq. (3.7d) with the expectation in (3.7d) written out explicitly as an integral. Similarly Eqs. (3.7b) and (5.2) (evaluated at $k = -1$) are alternative representations for $F_Y(y)$ as a functional of F_O.

REMARK 5.1. We refer to the RHS of (5.2) as the selection bias G-computation algorithm formula. Under semiparametric model **b** of Theorem (3.2), if the $q_k\left(\overline{\ell}_k, \underline{y}_{k+1}\right) \equiv 0$ for all k so that there is no non-ignorable selection bias for Y, then $j\left(\overline{\ell}_k, \underline{y}_{k+1}\right) = 1$ for all k and we obtain the standard g-computation algorithm formula of Robins (1986).

REMARK 5.2. Scharfstein, Rotnitzky, and Robins (1999) obtain a selection bias continuous time g-computation algorithm formula by explicitly writing out the expectation in Eq. (3.11) in terms of the joint distribution of the observables under semiparametric model **b** of Theorem 3.3 in the special case where the $\overline{L}(u)$ process jumps only at a finite number of non-random times. Richard Gill has generalized this result by allowing for $\overline{L}(u)$ processes that can jump in continuous time.

5.2. Selection odds model. Consider semiparametric model **b** for (Δ, L) of Theorem 3.2 with Y the outcome of interest. The odds ratio function of Y and Δ, $OR_{Y\Delta}(y)$ is defined to be $OR_{Y\Delta}(y) = \{pr[\Delta = 1 \mid Y = y]/pr[\Delta = 0 \mid Y = y]\}/\{pr[\Delta = 1 \mid Y = 0]/pr[\Delta = 0 \mid Y = 0]\} = \{f_Y[y \mid \Delta = 1]/f_Y[y \mid \Delta = 0]\}/\{f_Y[0 \mid \Delta = 1]/f_Y[0 \mid \Delta = 0]\}$. $OR_{Y\Delta}(y)$ is the unique functional of the distribution of $F_{Y,\Delta}$ that is a functional of both the conditional distribution of Y given Δ and of the distribution of Δ given Y. We refer to the semiparametric model **b** as a non-parametric selection odds model if we choose $\Phi(x)$ to be the logistic function $e^x/(1+e^x)$ since then $\ln OR_{Y\Delta}(y) = q(y)$. In the literature, a "selection" model is a model for selection bias that models the law of Δ given Y, and a pattern mixture model is a model for the law of Y given Δ. We see that our non-parametric selection odds model is the unique model that has both a pattern mixture and selection model interpretation and thus is of some independent interest. It is also of interest because, in general, physicians and other investigators have a fairly good intuitive sense of the meaning of the function $q(y)$ in a selection odds model because of the two equivalent useful ways to think about the meaning of $q(y)$ as encoded in the above equivalent definitions of $OR_{Y\Delta}(y)$.

We will now extend our study of selection odds models to the semiparametric model **b** of Theorem 3.2 for monotone missing data in discrete time with $\Phi(x)$ the logistic function. In this model, the chosen functions $q_k(\overline{L}_k, \underline{Y}_{k+1})$ represent the log odds ratio for drop-out at k. In the following, as discussed in Remark 2.10, we standardize our choice of q_k such that $q_k(\overline{L}_k, \underline{Y}_{k+1}) = 0$ if $\underline{Y}_{k+1} \equiv 0$. An alternative representation of our selection odds model is given in the following.

LEMMA 5.1. *Eq. (3.5) is true with $\Phi(x)$ logistic if and only if for $k = 0, \ldots, K$ $f_{\underline{Y}_{k+1}}(\underline{y}_{k+1} \mid \overline{L}_k, C = k) = f_{\underline{Y}_{k+1}}(\underline{y}_{k+1} \mid \overline{L}_k, C > k) \exp\{-q_k(\overline{L}_k, \underline{y}_{k+1})\}/c_k(\overline{L}_k)$ where*

$$(5.4)\quad c_k(\overline{L}_k) \equiv \int f_{\underline{Y}_{k+1}}\left(\underline{y}_{k+1} \mid \overline{L}_k, C > k\right) \exp\left\{-q_k\left(\overline{L}_k, \underline{y}_{k+1}\right)\right\} d\mu\left(\underline{y}_{k+1}\right).$$

We will now use the following lemma.

LEMMA 5.2. $B_K(\underline{y}_{K+1}) = f[\underline{y}_{K+1} \mid \overline{L}_K, C > K]$, $B_k^*(\underline{y}_{k+1}) = B_k(\underline{y}_{k+1})\left[\{1 - \Lambda_k^*\} + \Lambda_k^* e^{-q_k(\overline{L}_k, \underline{y}_{k+1})}/c_k(\overline{L}_k)\right]$, *and* $B_{m-1}(\underline{y}_m) = \int b_m^*(\underline{y}_{m+1}, \{\overline{L}_{m-1}, \ell_m = (v_m, y_m)\}) f(y_m \mid \overline{L}_{m-1}, C \geq m) dF[v_m \mid \overline{L}_{m-1}, Y_m = y_m, C \geq m]$.

REMARK 5.3. Note if $q_k(\overline{L}_k, \underline{y}_{k+1}) \equiv 0$ for all k so that there is no selection bias for Y, then $c_k(\overline{L}_k) \equiv 1$ and $B_k(\underline{y}_{k+1}) = B_k^*(\underline{y}_{k+1})$. For a selection odds model, $j(\overline{L}_k, \underline{y}_{k+1})$ has a particularly nice form. Specifically, if, in Eq. (3.5), $\Phi(x)$ is logistic, then

$$j\left(\overline{L}_m, \underline{y}_{m+1}\right) \equiv b_m^*\left(\underline{y}_{m+1}, \overline{L}_m\right) / b_m\left(\underline{y}_{m+1}, \overline{L}_m\right)$$
$$= \left(1 - \Lambda_m^*\right) + \Lambda_m^* \exp\left\{-q_k\left(\overline{L}_m, \underline{y}_{m+1}\right)\right\} / c_m\left(\overline{L}_m\right)$$

where the $c_m\left(\overline{\ell}_m\right)$ are obtained recursively starting with $m = K$ from Eq. (5.4).

6. Identification in causal inference problems.

6.1. The data and counterfactual data. Consider a study where we observe n i.i.d. copies of data $O = \left(\overline{A}(\tau), \overline{L}(\tau)\right)$, where τ is an administrative end of follow-up time, $\overline{A}(\tau)$ is a treatment process, $\overline{L}(\tau)$ is an outcome or response process and, for any $Z(u), \overline{Z}(t) \equiv \{Z(u); 0 \leq u \leq t\}$. We assume τ 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 $\overline{A} = \{A(u); 0 \leq u < \infty\}$ with $A(u)$ taking values in a set $\mathcal{A}(u)$ and the existence of underlying counterfactual random variables

$$\left\{\overline{L}_{\overline{a}};\ \overline{a} \in \overline{\mathcal{A}}\right\} \tag{6.1}$$

where $\overline{L}_{\overline{a}} = \{L_{\overline{a}}(u); 0 \leq u < \infty\}$, $\overline{a} = a(\cdot) = \left\{a(t); 0 \leq t < \infty \text{ and } a(t) \in \mathcal{A}(t)\right\}$ is a treatment plan (equivalently, regime or function) lying in a set of functions $\overline{\mathcal{A}}$. Given a regime $\overline{a}$, let $\overline{L}_{\overline{a}(u),0}$ 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:

$$\overline{L}_{\overline{a}(u),0}(u) = \overline{L}_{\overline{a}(t),0}(u) = \overline{L}_{\overline{a}}(u) = \overline{L}_{\overline{a}^\dagger}(u) \tag{6.2}$$

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

$$\overline{L}(\tau) = \overline{L}_{\overline{A}(\tau),0}(\tau)\ . \tag{6.3}$$

Eq. (6.3) 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. Robins (1987, 1997b) considers the sequential randomization (i.e., ignorable treatment assignment) assumption that for all t and $\overline{a} \in \overline{\mathcal{A}}$,

$$\underline{Y}_{\overline{a}}(t) \coprod A(t) \mid \overline{L}\left(t^-\right), \overline{A}\left(t^-\right) \tag{6.4}$$

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 (6.4) as the assumption of no unmeasured

confounders given prognostic factors $L(t)$. Because of measurability issues, (6.4) 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., $\overline{L}(t^-) \equiv \lim_{u \uparrow t} \overline{L}(u)$, (6.4) is formally, for each t_k,

$$(6.5) \quad f\left[A(t_k) \mid \overline{L}(t_k^-), \overline{A}(t_k^-), \underline{Y}_{\overline{a}}(t_k)\right] = f\left[A(t_k) \mid \overline{L}(t_k^-), \overline{A}(t_k^-)\right] .$$

where $f(\cdot \mid \cdot)$ is the conditional density of $A(t_k)$ with respect to a dominating measure $\mu(\cdot)$. 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. (6.4) is formally that

$$(6.6a) \qquad \lambda_A\left[t \mid \overline{L}(t^-), \overline{A}(t^-), \underline{Y}_{\overline{a}}(t)\right] = \lambda_A\left[t \mid \overline{L}(t^-), \overline{A}(t^-)\right]$$

and

$$(6.6b) \qquad \begin{aligned} f\left[A(t) \mid \overline{L}(t^-), \overline{A}(t^-), A(t) \neq A(t^-), \underline{Y}_{\overline{a}}(t)\right] = \\ f\left[A(t) \mid \overline{L}(t^-), \overline{A}(t^-), A(t) \neq A(t^-)\right] . \end{aligned}$$

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. (6.6a) 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. (6.6b) 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. Given (6.4), Robins (1987) shows that the marginal distribution of $Y_{\overline{a}}$ is identified by the g-computation algorithm formula, as discussed further below.

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

$$(6.7) \qquad f\left[\overline{A}(\tau) \mid \left\{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\right\}\right] \textit{ depends only on } O = \left(\overline{A}(\tau), \overline{L}(\tau)\right) .$$

Note that we can use ideas from the "missing data" literature because one's treatment history $\overline{A}(\tau)$ 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). We shall make the following non-identifiable assumption concerning the statistical models for the full data $\left(A, \left\{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\right\}\right)$ considered in this section. Given $\overline{a}_1$ and $\overline{a}_2$, let u_{12} be the smallest time u with $a_1(u) \neq a_2(u)$. Thus consider the following non-identifiable assumption.

ASSUMPTION A. For all $\overline{a}_1$ and $\overline{a}_2$ the conditional distribution of $\left(\overline{L}_{\overline{a}_1}, \overline{L}_{\overline{a}_2}\right)$ given $\overline{L}_{\overline{a}_1}(u_{12}^-)$ is non-degenerate.

LEMMA 6.1. *If Assumption A and CAR hold, then so does sequential randomization (6.4).*

Proof. Ignoring measured theoretic subtleties, we can assume without loss of generality that the $A(t)$ process jumps only at $t = 0, A(t) \in \{0, 1\}$, the $L_{\bar{a}} = Y_{\bar{a}}$ process jumps only a $t = 1$, and that (6.4) is false because

$$f[A(0) = 1 \mid Y_1(1)] = q[Y_1(1)] \ .$$

Although the last display does not violate the CAR assumption (6.7), nonetheless, it also implies $f[A(0) = 0 \mid Y_1(1)] = 1 - q[Y_1(1)]$ which does violate (6.7) unless $Y_1(1) = Y_0(1)$ w.p.1, which is prohibited by Assumption A. □

Lemma 6.1 has the following obvious partial converse if we strengthen (6.4).

LEMMA 6.2. *Suppose that (6.4) holds with* $\{\overline{L}_{\bar{a}}; \bar{a} \in \overline{\mathcal{A}}\}$ *replacing* $\underline{Y}_{\bar{a}}(t)$. *Then CAR holds.*

However, if (6.4) is not so strengthened, then, even under Assumption A, the converse to Lemma 6.1 is not true. Specifically, Robins (1997b, p. 83) gives examples where one would expect (6.4) to be true even when (6.7) is false. However, if (6.7) is the sole restriction imposed, this essentially places no restrictions on the joint distribution of the observable random variables O (Gill, van der Laan, Robins, 1997) and, thus, is not subject to empirical test. Thus, once (6.4) is assumed, we can impose (6.7) without affecting our (non-parametric) inference. In the following remark, we show by counterexample that without Assumption A, CAR (i.e., 6.7) does not imply sequential randomization (i.e., 6.4), in which case the g-computation algorithm formula cannot be used to compute the marginal distribution of $\overline{Y}_{\bar{a}}$.

REMARK 6.1. Suppose that $A(t)$ process jumps only at time $1^-, 2^-$ and $A(t)$ is a dichotomous $(0, 1)$ variable. Let $Y_{ij} = (Y_{ij}(1), Y_{ij}(2))$ be $Y_{\bar{a}=(i,j)}$ [i.e., $\overline{Y}_{\bar{a}=(a(1^-),a(2^-))}$ with $a(1^-) = i$ and $a(2^-) = j$]. Suppose, in violation of Assumption A, that $Y_{01}(2) = Y_{11}(2)$ with probability 1. That is, $a(1^-)$ has no direct effect on Y at time 2 when $a(2^-)$ is set to 1. Further suppose: $Y_{0j}(1) = Y_{i0}(2) = 0$ with probability 1 for all i and j. That is, Y is zero at time 1 or 2 if one receives treatment level 0 at times 1^- or 2^-, respectively. For notational convenience, write $A(1^-)$ and $A(2^-)$ as A_1 and A_2 respectively. Finally assume $Y_{10}(1)$ and $Y_{01}(2)$ are highly correlated and that

$$(6.8a) \qquad pr[A_1 = 1, A_2 = 0 \mid \{Y_{ij}; i, j = 1, 2\}] = \frac{1}{8} + \left(\frac{1}{8}\right) Y_{10}(1)$$

$$(6.8b) \qquad pr[A_1 = 0, A_2 = 1 \mid \{Y_{ij}; i, j = 1, 2\}] = \frac{1}{8} + \left(\frac{1}{8}\right) Y_{01}(2)$$

and

$$(6.8c) \quad pr[A_1 = 1, A_2 = 1 \mid \{Y_{ij}; i, j = 1, 2\}] = \frac{3}{4} - \frac{1}{8} Y_{10}(1) - \frac{1}{8} Y_{01}(2) \ .$$

Now, by (6.2), $Y_{10}(1) = Y_{11}(1)$ w.p.1 and, by assumption, $Y_{01}(2) = Y_{11}(2)$. Thus one can substitute Y_{11} for Y_{10} and Y_{01} in (6.8c) and check that the data are CAR. However, we now show that $pr[A_1 = 0 \mid Y_{10}(1)] \neq pr[A_1 = 0]$ in violation of (6.4). Specifically, $pr[A_1 = 0 \mid Y_{01}(2)]$ depends on $Y_{01}(2)$ by (6.8b). Furthermore, $pr[A_1 = 0 \mid Y_{10}(1)] = pr[A_1 = 0, A_2 = 1 \mid Y_{10}(1)] = 1/8 + (1/8)E[Y_{01}(2) \mid Y_{10}(1)]$ which depends, by the correlation assumption, on $Y_{10}(1)$.

This example was derived as follows. There are underlying dichotomous variables $Y^{(1)}, Y^{(2)}$. Furthermore, $Y_{10}(1) \equiv Y_{11}(1) \equiv Y^{(1)}$ and $Y_{01}(2) \equiv Y_{11}(2) \equiv Y^{(2)}$. Also $Y_{0i}(1) = Y_{i0}(2) = 0$ for $i \in \{1, 2\}$. We observe $\left(A(1^-), A(1^-)Y^{(1)}, A(2^-), A(2^-)Y^{(2)}\right)$ with the CAR probabilities given above. Under the CAR assumption, Gill, van der Laan, and Robins (1997) show that the joint distribution of $\{Y_{ij}; i, j = 1, 2\}$ is identified but not by the g-computation algorithm formula.

REMARK 6.2. Assumptions concerning the joint distribution of $\left(\overline{L}_{\overline{a}_1}, \overline{L}_{\overline{a}_2}\right)$ given $\overline{L}_{\overline{a}_1}\left(u_{12}^-\right)$ plus the assumption that the data are CAR place no restriction on the joint distribution of the observed data O. However, as the above example shows, such assumptions may be sufficient to rule out sequential randomization. Indeed, in the example of Remark 6.1, the assumption that $Y_{01}(2) = Y_{11}(2)$ w.p.1 alone is sufficient to rule out the sequential randomization assumption, since the two assumptions together imply the restriction on the joint distribution of the observed data that $\Omega(j) \equiv \int E[Y(2) \mid A_1 = j, A_2 = 1, Y(1)]dF[Y(1) \mid A_1 = j]$ is not a function of j. However, assuming both CAR and that Assumption A is violated is not sufficient to conclude that sequential randomization is false. To see this, consider the example of Remark 6.1 but assume that the probability of the event $A_1 = A_2 = 1$ was zero. Then it is easy to check that CAR is equivalent to sequential randomization even though Assumption A is assumed false.

REMARK 6.3. The example of Remark 6.1 can be viewed as a discrete-time version of interval censored data in which we assume there is an underlying failure time variable T and we define $Y^{(1)} = I(T \leq 1), Y^{(2)} = I(T \leq 2)$ and $A_j = 1$ if a subject was monitored at time j. On the other hand, when the probability of the event $A_1 = A_2 = 1$ is zero, we can view the example as a discrete-time version of current status data in which each subject is monitored only once. We can then conclude from our previous discussion that if we wish to estimate the distribution of our failure time random variable T under the sole assumption that the data are CAR, the distribution of T can be obtained using the g-computation algorithm formula in the case of current status data but cannot be so obtained in the case of interval censored data. This fact underlies the observation that the efficient score for estimating functionals of the distribution of T has an elegant closed form martingale representation in the case of current status data but not in the case of interval censored data (van der Laan and Robins, 1998).

We now consider identification of the law of $Y_{\bar{a}}$ when (6.4) is false due to confounding by unmeasured factors.

6.2. Identification with unmeasured confounders. Suppose the A-process jumps only at fixed times $0, 1, 2, \ldots, K$ and the L-process jumps only at times $0^-, 1^-, \ldots, K+1^-$. Write, for notational convenience, $A_k = A(k)$ and $L_k = L(k^-)$. $\overline{A}_k \equiv (A_0, \ldots, A_k)\underline{Y}_{\bar{a},k} = (Y_{\bar{a},k}, Y_{\bar{a},k+1}, \ldots, Y_{\bar{a},K+1})$. We then have the following theorem.

THEOREM 6.1. *Suppose that A_k is discrete and*

$$f_{A_k}\left[a_k \mid \overline{A}_{k-1} = \overline{a}_{k-1}, \overline{L}_k\right] > 0 \ w.p.1 \tag{6.9}$$

for all $\overline{a}_k \in \overline{\mathcal{A}}_k, k = 0, \ldots, K$. Then, the semiparametric model **b** *for $\left(\overline{A}, \{\overline{L}_{\bar{a}}; \bar{a} \in \overline{\mathcal{A}}\}\right)$ characterized by the sole restriction*

$$\begin{aligned} &f_{A_k}\left[a_k \mid \overline{A}_{k-1} = \overline{a}_{k-1}, \overline{L}_k, \underline{Y}_{\bar{a},k+1}\right] \\ &\qquad = 1 - \Phi\left[h_k\left(\overline{L}_k, \bar{a}\right) + q_k\left(\overline{L}_k, \underline{Y}_{\bar{a},k+1}, \bar{a}\right)\right] \end{aligned} \tag{6.10}$$

with (i) $\Phi(x)$ a known continuous, monotone increasing, distribution function, (ii) $q_k\left(\overline{\ell}_k, \underline{y}_{k+1}, \bar{a}\right)$ a known function, and (iii) $h_k\left(\overline{\ell}_k, \bar{a}\right)$ unknown is a non-parametric model for the law F_O of $O = \left(\overline{A}(\tau), \overline{L}(\tau)\right)$ and the distributions $B_m\left(\underline{y}_{m+1}, \bar{a}\right) \equiv b_m\left(\underline{y}_{m+1}, \overline{L}_m, \bar{a}\right) \equiv f_{\underline{Y}_{\bar{a},m+1}}\left(\underline{y}_{m+1} \mid \overline{L}_m, \overline{A}_m = \overline{a}_m\right)$, $B^*_m\left(\underline{y}_{m+1}, \bar{a}\right) \equiv b^*_m\left(\underline{y}_{m+1}, \overline{L}_m, \bar{a}\right) \equiv f_{\underline{Y}_{\bar{a},m+1}}\left(\underline{y}_{m+1} \mid \overline{L}_m, \overline{A}_m = \overline{a}_{m-1}\right)$, and $f_{Y_{\bar{a}}}(y)$ and the functions $h_k\left(\overline{L}_k, \bar{a}\right)$ are identified from data on O. Specifically, suppress the dependence on $\bar{a}$ in the notation and write, for a given $\bar{a}$, $h_k\left(\overline{\ell}_k\right) \equiv h_k\left(\overline{\ell}_k, \bar{a}\right), q_k\left(\overline{\ell}_k, \underline{y}_{k+1}\right) \equiv q_k\left(\overline{\ell}_k, \underline{y}_{k+1}, \bar{a}\right)$, etc., and define C to be the minimum of $K+1$ and the first time for which $A_k \neq a_k$. Write $\Delta = I\left(C = K+1\right)$. Then, with Λ_k defined as in (3.5), $h_m\left(\overline{L}_m\right)$ is the unique solution to (3.7a) and $F_{Y_{\bar{a}}}(y)$ and $F_{\underline{Y}_{\bar{a},m+1}}\left(\underline{y}_{m+1} \mid \overline{L}_m, \overline{A}_{m-1} = \overline{a}_{m-1}\right)$, $F_{\underline{Y}_{\bar{a},m+1}}\left(\underline{y}_{m+1} \mid \overline{L}_m, \overline{A}_m = \overline{a}_m\right)$ are given by the RHS of (3.7b), (3.7c), and (3.7d) respectively. Further, $f_{\underline{Y}_{\bar{a},k+1}}\left(\underline{y}_{k+1} \mid \overline{\ell}_k, \overline{A}_k = \overline{a}_k\right)$ is given by the selection bias g-computation algorithm formula [i.e., the RHS of (5.2)] with $j(\cdot,\cdot)$ given by (5.3). In particular, if $\Phi(x)$ is logistic, $j(\cdot,\cdot)$ is given by (5.5). Note this implies that

$$c_k(\overline{L}_k) \equiv \int f_{\underline{Y}_{\bar{a},k+1}}(\underline{y}_{k+1} \mid \overline{L}_k, \overline{A}_k = \overline{a}_k) \exp[-q_k(\overline{L}_k, \underline{y}_{k+1})] d\mu(\underline{y}_{k+1}) . \tag{6.11}$$

REMARK 6.4. Although we do not give the proof of Theorem 6.1, we do here address an important issue. Although true, it is not obvious that given any F_O, there exists a law for $\left(\overline{A}, \{\overline{L}_{\bar{a}}; \bar{a} \in \overline{\mathcal{A}}\}\right)$ satisfying (6.10). We take the simplest possible setting to make clear why indeed there is such a law. The general setting follows by arguing recursively. Consider the case

where $K = 0$, $L_1 = Y_1$, and $L_0 = \emptyset$ and define $A = A_0$ and $Y = Y_1$. Then (6.10) says

$$f_A[a \mid Y_a] = 1 - \Phi[h(a) + q(Y_a, a)]$$

which, by Bayes' Theorem, is equivalent to

$$\begin{aligned} f_{Y_a}(y \mid A \neq a) = f_{Y_a}(y \mid A = a)\{pr(A \neq a)/pr(A = a)\} \\ \{[1 - \Phi(h(a) + q(y, a))]/\Phi(h(a) + q(y, a))\} \end{aligned} \tag{6.12a}$$

where $h(a)$ is the unique solution to

$$\begin{aligned} pr(A = a)/pr(A \neq a) = \\ \int f_{Y_a}(y \mid A = a)\{[1 - \Phi(h(a) + q(y, a))]/\Phi(h(a) + q(y, a))\}\, d\mu(y). \end{aligned} \tag{6.12b}$$

Note, by Φ being a strictly increasing distribution function and assuming as we do throughout that the integral is finite, (6.12b) is guaranteed to have a unique solution $h(a)$. Since the RHS of (6.12a) thus only depends on the law F_O of the observed data and the known function $q(y, a)$, it is obvious that we can generate a joint distribution for $(A, \{Y_a; a \in \mathcal{A}\})$ satisfying (6.12a).

7. Arbitrary continuous or discrete treatments.

7.1. Selection odds models for discrete or continuous treatments. We will generalize the selection odds model of Section 6 by allowing A_k to be discrete or continuous. Write $\Pi_k^*(\bar{a}_k) = \pi_k^*(\bar{L}_k, \bar{a}_k) = f_{A_k}(a_k \mid \bar{L}_k, \bar{A}_{k-1} = \bar{a}_{k-1})$ and $\Pi_k(\underline{y}_{k+1}, \bar{a}) \equiv \pi_k(\bar{L}_k, \underline{y}_{k+1}, \bar{a}) = f_{A_k}(a_k \mid \bar{L}_k, \bar{A}_{k-1} = \bar{a}_{k-1}, \underline{Y}_{\bar{a},k+1} = \underline{y}_{k+1})$. Again define $B_k(\underline{y}_{k+1}, \bar{a}) \equiv b_k(\underline{y}_{k+1}, \bar{L}_k, \bar{a}) = f_{\underline{Y}_{\bar{a},k+1}}(\underline{y}_{k+1} \mid \bar{L}_k, \bar{A}_k = \bar{a}_k)$, $B_k^*(\underline{y}_{k+1}, \bar{a}) = b_k^*(\underline{y}_{k+1}, \bar{L}_k, \bar{a}) = f_{\underline{y}_{\bar{a},k+1}}(\underline{y}_{k+1} \mid \bar{L}_k, \bar{A}_{k-1} = \bar{a}_{k-1})$. Let $J_k(\underline{y}_{k+1}, \bar{a}) \equiv j_k(\bar{L}_k, \underline{y}_{k+1}, \bar{a}) \equiv \Pi_k^*(\bar{a}_k)/\Pi_k(\underline{y}_{k+1}, \bar{a}) = B_k^*(\underline{y}_{k+1}, \bar{a})/B_k(\underline{y}_{k+1}, \bar{a})$ where the last equality is by Bayes' theorem. Consider now the selection-bias g-computation algorithm formula identity

$$\begin{aligned} b_k\left(\underline{y}_{k+1}, \bar{\ell}_k, \bar{a}\right) &= \int \cdots \iint \prod_{m=k+1}^{K+1} f_{Y_m}\left[y_m \mid \bar{\ell}_{m-1}, \bar{a}_{m-1}\right] dF\left[v_m \mid \right. \\ &\qquad \left. y_m, \bar{\ell}_{m-1}, \bar{a}_{m-1}\right] \prod_{m=k+1}^{K} j_m\left(\bar{\ell}_m, \underline{y}_{m+1}, \bar{a}\right) \\ &= \int \cdots \iint f\left[\underline{y}_{k+1}, \underline{v}_{k+1}, \underline{a}_{k+1} \mid \right. \\ &\qquad \left. \bar{\ell}_k, \bar{a}_k\right] \left\{\prod_{m=k+1}^{K} \pi_m\left(\bar{\ell}_m, \underline{y}_{m+1}, \bar{a}\right)\right\}^{-1} \prod_{m=k+1}^{K+1} d\mu(v_m) \end{aligned} \tag{7.1a}$$

and

$$b_k^*\left(\underline{y}_{k+1}, \bar{\ell}_k, \bar{a}\right) = \int \cdots \iint \prod_{m=k+1}^{K+1} f_{Y_m}\left[y_m \mid \bar{\ell}_{m-1}, \bar{a}_{m-1}\right] dF\left[v_m \mid y_m, \bar{\ell}_{m-1}, \bar{a}_{m-1}\right] \prod_{m=k}^{K} j_m\left(\bar{\ell}_m, \underline{y}_{m+1}, \bar{a}\right)$$

(7.1b)

$$= \int \cdots \iint f\left[\underline{y}_{k+1}, \underline{v}_{k+1}, \underline{a}_{k+1} \mid \bar{\ell}_k, \bar{a}_k\right] \left\{\prod_{m=k}^{K} \pi_m\left(\bar{\ell}_m, \underline{y}_{m+1}, \bar{a}\right)\right\}^{-1} \pi_k^*\left(\bar{\ell}_k, \bar{a}_k\right) \prod_{m=k+1}^{K+1} d\mu\left(v_m\right).$$

It follows from (7.1a) that to make $b_k\left(\underline{y}_{k+1}, \bar{\ell}_k, \bar{a}\right)$ identifiable from data on O, we need to make the $j_k\left(\bar{\ell}_k, \underline{y}_{k+1}, \bar{a}\right)$ or equivalently the $\pi_k\left(\bar{\ell}_k, \underline{y}_{k+1}, \bar{a}\right)$ identifiable. One approach to doing so is given in the following lemma.

LEMMA 7.1. *Consider the semiparametric selection odds model for the distribution of* $\left(\overline{A}, \left\{\overline{L}_{\bar{a}}; \bar{a} \in \overline{\mathcal{A}}\right\}\right)$ *that imposes the sole restriction that*

$$f_{\underline{Y}_{\bar{a},k+1}}\left(\underline{y}_{k+1} \mid \overline{L}_k, A_k \neq a_k, \overline{A}_{k-1} = \bar{a}_{k-1}\right) = C_k\left(\bar{a}\right) B_k\left(\underline{y}_{k+1}, \bar{a}\right) Q_k^*\left(\underline{y}_{k+1}, \bar{a}\right) \tag{7.2}$$

where (i) $Q_k^*\left(\underline{y}_{k+1}, \bar{a}\right) \equiv q_k^*\left(\overline{L}_k, \underline{y}_{k+1}, \bar{a}\right)$ *is a known non-negative function, and (ii)* $C_k\left(\bar{a}\right) = c_k\left(\overline{L}_k, \bar{a}\right)$ *is a normalizing constant chosen to make the LHS of (7.2) a density. Then, if (6.9) holds, the* $C_k\left(\bar{a}\right)$ *and the* $J_k\left(\underline{y}_{k+1}, \bar{a}\right)$ *are identified.*

In particular, it follows immediately from its definition that

$$J_k\left(\underline{y}_{k+1}, \bar{a}\right) = \Pi_k^*\left(\bar{a}_k\right) + \left\{1 - \Pi_k^*\left(\bar{a}_k\right)\right\} Q_k^*\left(\underline{y}_{k+1}, \bar{a}\right) \left\{C_k\left(\bar{a}\right)\right\}^{-1} \tag{7.3a}$$

if

$$pr\left[A_k = a_k \mid \overline{L}_k, \overline{A}_{k-1} = \bar{a}_{k-1}\right] \neq 0, \tag{7.3b}$$

and

$$J_k\left(\underline{y}_{k+1}, \bar{a}\right) = Q_k^*\left(\underline{y}_{k+1}, \bar{a}\right) \left\{C_k\left(\bar{a}\right)\right\}^{-1}, \quad \textit{otherwise.} \tag{7.3c}$$

[For example, if A_k is continuous and was measured at each occasion k, we would expect (7.3b) to be false and $f_{a_k}\left(a_k \mid \overline{L}_k, \overline{A}_{k-1} = \bar{a}_{k-1}\right)$ to be a density with respect to Lesbesgue measure.] Then all necessary quantities can be calculated from the distribution of the observed data O by the following backward recursion. First, $B_K\left(\underline{y}_{K+1}, \bar{a}\right) = f_{\underline{Y}_{K+1}}\left(\underline{y}_{K+1} \mid \overline{L}_K, \overline{A}_K =\right.$

$\overline{a}_K)$. Then, given $B_k(\underline{y}_{k+1}, \overline{a})$, we can calculate $C_k(\overline{a})$, $J_k(\underline{y}_{k+1}, \overline{a})$, and $B_{k-1}(\underline{y}_k, \overline{a})$ as follows. By its definition as a normalizing constant, we calculate $C_k(\overline{a}) = \int B_k(\underline{y}_{k+1}, \overline{a}) Q_k^*(\underline{y}_{k+1}, \overline{a}) d\mu(\underline{y}_{k+1})$. We then calculate $J_k(\underline{y}_{k+1}, \overline{a})$ by (7.3a)–(7.3c). Finally, we calculate $B_{k-1}(\underline{y}_k, \overline{a})$ from (7.1).

REMARK 7.1. If $Q_k^*(\underline{y}_{k+1}, \overline{a})$ does not depend on $\underline{y}_{k+1}$ so that (5.4) holds, then the $j_k(\overline{\ell}_k, \underline{y}_{k+1}, \overline{a})$ are all identically 1 and (7.1) reduces to the usual g-computation algorithm formula.

REMARK 7.2. A selection odds model is a non-parametric model for F_O in the sense that even though $Q_k^*(\underline{y}_{k+1}, \overline{a})$ is a known function, the model is compatible with any law F_O of O since the restriction (7.2) is not identifiable from data O. It is interesting to note that if there is selection bias due to unmeasured confounding (i.e., $Q_k^*(\underline{y}_{k+1}, \overline{a})$ depends on $\underline{y}_{k+1}$), then the conditional densities $b_k(\underline{y}_{k+1}, \overline{\ell}_k, \overline{a})$ of the counterfactuals do not depend on the densities of the treatment process $f(a_k \mid \overline{\ell}_k, \overline{a}_{k-1})$ if and only if (7.3b) is false with probability 1 for each $\overline{a}_k$.

REMARK 7.3. We can restate restriction (7.2) in a manner closely related to (6.10). Indeed, its correspondence is exact when (7.3b) is true w.p.1. Specifically, the NPI model defined by restriction (7.2) is equivalent to the following NPI model:

(i) If (7.3b) is true,

$$\pi_k\left(\overline{L}_k, \underline{y}_{k+1}, \overline{a}\right) = 1 - expit\left[h_k\left(\overline{L}_k, \overline{a}\right) + q_k\left(\overline{L}_k, \underline{y}_{k+1}, \overline{a}\right)\right] \tag{7.4}$$

with $h_k(\cdot, \cdot)$ unrestricted and $q_k(\cdot, \cdot, \cdot)$ known. Specifically, the models are related by

$$h_k\left(\overline{L}_k, \overline{a}\right) = \ln\left[\left\{\Pi_k^*\left(\overline{a}_k\right)\right\}^{-1}\left\{1 - \Pi_k^*\left(\overline{a}_k\right)\right\} C_k\left(\overline{a}\right)^{-1}\right] \tag{7.5}$$

and

$$q_k\left(\overline{L}_k, \underline{y}_{k+1}, \overline{a}\right) = \ln Q_k^*\left(\underline{y}_{k+1}, \overline{a}\right) . \tag{7.6}$$

(ii) If (7.3b) is false,

$$\Pi_k\left(\underline{y}_{k+1}, \overline{a}\right) = \exp\left[-\left\{h_k\left(\overline{L}_k, \overline{a}\right) + q_k\left(\overline{L}_k, \underline{y}_{k+1}, \overline{a}\right)\right\}\right] \tag{7.7}$$

with $h_k(\cdot, \cdot)$ unrestricted and $q_k(\cdot, \cdot, \cdot)$ known. Specifically, the models are related by $q_k(\overline{L}_k, \underline{y}_{k+1}, \overline{a})$ being given by (7.6) and

$$h_k\left(\overline{L}_k, \overline{a}\right) = -\ln\left[\Pi_k^*\left(\overline{a}_k\right) C_k\left(\overline{a}\right)\right] . \tag{7.8}$$

Furthermore, whether or not (7.3b) is true, the $h_k(\overline{\ell}_k, \overline{a})$ and thus the $\pi_k(\overline{\ell}_k, \underline{y}_{k+1}, \overline{a})$ are identified from the law F_O of O and the known function $q_k^*(\overline{\ell}_k, \underline{y}_{k+1}, \overline{a})$ as the unique solution to the identity

$$E_O\left[w_k(\underline{A}_k)/\prod_{m=k}^{K}\pi_m(\overline{L}_m,\underline{Y}_{m+1},\overline{A})\mid \overline{L}_k=\overline{\ell}_k,\overline{A}_{k-1}=\overline{a}_{k-1}\right] = \int w_k(\underline{a}_k)\,d\mu(\underline{a}_k) \tag{7.9}$$

for all functions $w_k(\underline{a}_k), k = K, K-1, \ldots, 0$. Specifically, we proceed recursively and first identify $h_K(\overline{L}_K, \overline{a})$ by (7.9) with $k = K$. We then identify $h_{K-1}(\overline{L}_{K-1}, \overline{a})$ by (7.9) with $k = K-1$, etc.

Proof. That (7.9) is true follows from the fact that by (7.1b), (7.9) equals $\int\{\int b_k^*(\underline{y}_{k+1}, \overline{\ell}_k, \underline{a})d\mu(\underline{y}_{k+1})\}w_k(\underline{a}_k)\,d\mu(\underline{a}_k) = \int w_k(\underline{a}_k)\,d\mu(\underline{a}_k)$. Finally it is straightforward to show that the solution $h_k(\overline{\ell}_k, \overline{a})$ must be unique. □

By a similar argument, given the $\pi_k\left(\overline{\ell}_k, \underline{y}_{k+1}, \overline{a}\right)$, the $b_k\left(\underline{y}_{k+1}, \overline{\ell}_k, \overline{a}\right)$ are the unique solutions to the identity

$$E_O\left[w(\underline{Y}_{k+1}, \underline{A}_{k+1})/\prod_{m=k+1}^{K}\Pi_m(\underline{Y}_{m+1}, \overline{A})\mid \overline{L}_k=\overline{\ell}_k, \overline{A}_k=\overline{a}_k\right] = \int w_k\left(\underline{y}_{k+1}, \underline{a}_{k+1}\right) b_k\left(\underline{y}_{k+1}, \overline{\ell}_k, \overline{a}\right) d\mu(\underline{a}_{k+1})\,d\mu\left(\underline{y}_{k+1}\right) \tag{7.10}$$

for all functions $w\left(\underline{y}_{k+1}, \underline{a}_{k+1}\right)$. We thus have the following theorem, which is almost a perfect analog of Theorem 3.2.

THEOREM 7.1. *The semiparametric model* **b** *for* $(\overline{a}, \{\overline{L}_a; \overline{a} \in \overline{\mathcal{A}}\})$ *characterized by the sole restrictions (7.4) and (7.7) with the* $q_k(\cdot,\cdot,\cdot)$ *known functions and the* $h_k(\cdot,\cdot)$ *completely unknown, is a non-parametric model for the law* F_O *for the observed data* $O = (\overline{A}(C), \overline{L}(C))$. *Furthermore, if (6.9) holds, the* $h_k(\overline{\ell}_k, \overline{a})$ *and* $b_k\left(\underline{y}_{k+1}, \overline{\ell}_k, \overline{a}\right)$ *are identified from the law of* O *as the unique solutions to (7.9) and (7.10). Further, the unique solution to (7.10) is Eq. (7.1a) with* $j_k\left(\overline{\ell}_k, \underline{y}_{k+1}, \overline{a}\right)$ *given by (7.3).*

7.2. Difficulty with semiparametric inference selection odds models in high-dimensional problems.

7.2.1. Difficulty with modeling $h_k(\cdot,\cdot)$. When the covariates L_k are high-dimensional with continuous components, we might hope, as in Sec. 4, to respond to the curse of dimensionality by modeling the functions $h_k(\overline{\ell}_k, \overline{a})$ in the representation (7.4) or (7.7) of our NPI selection odds model. Unfortunately, this approach fails.

To see why, consider the simplest setting where $K = 0$, $Y_1 = L_1$ and A_0 are dichotomous $(0,1)$ variables. Write $Y \equiv Y_1, L = L_0, A = A_0$. Then our NPI selection odds model defined by restriction (7.4) can be written for $a \in \{0,1\}$

$$pr[A=a \mid L, Y_a=y] = \pi(L,y,a) = 1 - expit[h(L,a)+q(L,y,a)] \tag{7.11}$$

$$q(L,y,a) \text{ known} \tag{7.12}$$

and $h(L,a)$ completely unrestricted. Now suppose L is high-dimensional with continuous components. Suppose that the contrast $E(Y_1) - E(Y_0)$ is the parameter of interest. We might hope, in analogy with our approach in Sec. 4, to estimate our contrast of interest using semiparametric augmented inverse probability of treatment weighted estimators as follows. First, in analogy to Eq. (4.1), we impose the additional restriction that the function $h(\ell,a)$ lies in a parametric family, i.e., for $a \in \{0,1\}$

$$h(\ell,a) = h(\ell,a;\gamma_a^*) \tag{7.13}$$

where $h(\cdot,\cdot,\cdot)$ is a known function and γ_0^* and γ_1^* are unknown parameter vectors to be estimated. Then, by Eq. (7.9), for $a \in \{0,1\}$, we estimate γ_a^* by $\widehat{\gamma}_a^*$ satisfying

$$0 = n^{-1} \sum_i \{I(A_i = a)/\pi(L_i, Y_i, a; \widehat{\gamma}_a) - 1\} w(a, L_i) \tag{7.14}$$

where $w(a, L_i)$ is a user-supplied vector function of the dimension of γ_a and $\pi(\ell, y, a; \widehat{\gamma}_a)$ is defined as in Eq. (7.11) with $h(\ell, a; \widehat{\gamma}_a)$ replacing $h(\ell, a)$. We then estimate $E(Y_a)$ by

$$\widehat{E}(Y_a) = n^{-1} \sum_i I(A_i = a) Y_i/\pi(L_i, Y_i, a; \widehat{\gamma}_a) \ . \tag{7.15}$$

The difficulty with this approach is that in general there will exist no joint distribution for (Y_1, Y_0, A, L) compatible with our estimates

$$\left(\widehat{E}(Y_1), \widehat{E}(Y_0), \widehat{\gamma}_0, \widehat{\gamma}_1\right) \ . \tag{7.16}$$

This is because $pr[A = 0 \mid L]$ is separately identifiable from data on L_i, A_i, $A_i Y_i, i = 1, \ldots, n$ and also identifiable from the data $L_i, A_i, (1 - A_i) Y_i, i = 1, \ldots, n$.

Specifically, in order that there exists a joint distribution, say $\widehat{F}$ for (Y_1, Y_0, A, L) compatible with our estimate Eq. (7.16), requires that there exist densities $\widehat{F}_{Y_a}(1 \mid \ell) \equiv \widehat{F}_{Y|A,L}(y \mid \ell, a)$ and $\widehat{f}(\ell)$ such that

$$\widehat{E}(Y_a) = \int \widehat{f}_{Y_a}(1 \mid \ell)\, d\widehat{F}(\ell), a \in \{0,1\} \tag{7.17}$$

for which the following two equations are equal with probability one.

$$\widehat{pr}[A = 0 \mid L] = \sum_{y=0}^{1} \pi(L, y, 0; \widehat{\gamma}_0)\, \widehat{f}_{Y_0}(y \mid L) \tag{7.18}$$

and

$$\widehat{pr}[A=0 \mid L] = 1 - \widehat{pr}[A=1 \mid L] = 1 - \sum_{y=0}^{1} \pi(L, y, 1; \widehat{\gamma}_1)\, \widehat{f}_{Y_1}(y \mid L) \ . \tag{7.19}$$

In general, there will not exist any joint law satisfying (7.17) and having (7.18) and (7.19) equal for all L.

There are several possible philosophical and/or practical views we might take of this inconvenient fact.

1. View the models $\{h(\ell, a; \gamma_a)\}$ as a "sieve" in which we allow the dimension of γ_a to increase with sample size n such that, as $n \to \infty$, the model becomes dense in all functions $h(\ell, a)$. We then content ourselves with the notion that as $n \to \infty$, our incompatibility problem disappears, since we know that if $h(\ell, a)$ unrestricted, our selection odds model is a non-parametric model for the law of the observed data $O = (Y, L, A)$. We simply accept that, with any finite sample size, we have an incompatible model. One could argue that this approach is related to similar approaches taken in other parts of statistics. For example, the Dabrowska estimator (Dabrowska, 1988) of a bivariate survival function for independently right-censored data is not a true survival function at sample size n, although, as $n \to \infty$, it converges to a survival function. Similarly, Edgeworth or higher order kernel approximations to densities are not densities for any fixed sample size n, since for certain values of x they may be negative. Nonetheless, for each fixed x, they become positive as $n \to \infty$.
2. A second approach is that we could try to modify our estimates (7.16) by replacing them with new estimates based on those obtained from the closest (in some metric) joint distribution for (Y_1, Y_0, L, A).

This second option we do not know how to implement. The first option seems quite unsatisfactory, especially as we do not even know how to pick a model for $h(\ell, a)$ so that we are even close to a joint distribution at finite sample sizes.

7.2.2. Difficulty with modeling $C_k(\overline{a})$. An alternative approach would be to still use the augmented inverse probability of treatment weighted estimators (7.14)–(7.15) to obtain the estimates (7.16), except that we now model directly the $C_k(\overline{a}) = c_k(\overline{L}_k, \overline{a})$ and the $\Pi^*(\overline{a}_k)$ rather than the $h_k(\overline{L}_k, \overline{a})$. We then estimate the $h_k(\overline{L}_k, \overline{a})$ using the identity (7.5). To fix ideas, again consider our simple model with $O = (A, L, Y)$ used in the previous subsection. Then it is unproblematic to specify and fit a parametric model $\pi(L, a; \eta)$ for $\Pi^*(a) \equiv \pi^*(L, a) \equiv pr[A = a \mid L]$ by maximizing $\prod_{i=1}^n \pi^*(L_i, A_i; \eta)$ with respect to η. We also specify a parametric model

$$c(L, a) = c(L, a; \gamma_a^*) \tag{7.20}$$

for $c(L, a)$. We then precede as in the last subsection, where now $\pi(\ell, y, a; \widehat{\gamma}_a)$ is again as defined in Eq. (7.11) but with $h(\ell, a; \widehat{\gamma}_a, \widehat{\eta})$ [via (7.5)] replacing $h(\ell, a)$ and $\widehat{\eta}$ suppressed in the notation. As argued in the last

subsection, in general, then there will exist no joint law for (Y_1, Y_0, A, L) consistent with the estimators (7.16) and $\widehat{\eta}$.

Thus, in general, our attempt to use augmented inverse probability of treatment weighted estimators to estimate our selection odds model has failed. There are two options. Either we keep the selection odds model but change our estimation procedure or we replace the selection odds model with another NPI model which allows a simple semiparametric inverse probability of treatment weighted estimator while avoiding incompatible models.

We shall consider both approaches in Sec. 8.

8. Sensitivity analysis for multivariate structural models. In this section, we discuss sensitivity analysis for both multivariate structural nested models and multivariate marginal structural models. In Sections 8.1–8.4, we consider structural nested models.

8.1. Structural nested models.

8.1.1. Structural nested distribution models (SNDMs). In this section, we suppose the outcome $\underline{Y}_m = (Y_m, \ldots, Y_{K+1})$ has a continuous multivariate distribution given $\overline{L}_{m-1}, \overline{A}_{m-1}$ with probability 1. Then we can do a sensitivity analysis based on g-estimation of multivariate structural nested distribution models. The foundations of this analysis are again formed by a class of non-parametric (just) identified (NPI) models which we shall call SNDMs.

Given a history $\overline{a}$, the history $(\overline{a}_m, 0)$ is the history $\overline{a}^*$ that agrees with $\overline{a}$ through time m and is zero thereafter, where zero is the baseline level of treatment. Following Robins et al. (1992, Appendix 2), we define the multivariate blip-down functions $\gamma_m(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m) = [\gamma_{m,m+1}(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m), \cdots, \gamma_{m,K+1}(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m)]$ as the unique solution to

$$\text{(8.1a)}\quad F_{\underline{Y}_{(\overline{a}_{m-1},0),m+1}|\overline{\ell}_m,\overline{a}_m}[\gamma_m(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m)] = F_{\underline{Y}_{(\overline{a}_m,0),m+1}||\overline{\ell}_m,\overline{a}_m}(\underline{y}_{m+1})$$

satisfying

$$\text{(8.1b)}\quad \gamma_{m,k}\left(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m\right) \textit{ is a function of } \overline{y}_{K+1} \textit{ only through } \overline{y}_k$$

for $m = 0, \ldots, K$. To be specific, define $z^{m:k} = (z_m, \ldots, z_k)$. Then

$$\gamma_{m,m+1}\left(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m\right) = F^{-1}_{Y_{(\overline{a}_{m-1},0),m+1}|\overline{\ell}_m,\overline{a}_m} \circ F_{Y_{(\overline{a}_m,0),m+1}|\overline{\ell}_m,\overline{a}_m}(y_{m+1}),$$

and

$$\gamma_{m,k}\left(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m\right) = F^{-1}_{Y_{(\overline{a}_{m-1},0),k}|\overline{\ell}_m,\overline{a}_m, Y^{m+1:k-1}_{(\overline{a}_{m-1},0)} = \gamma_m^{m+1:k-1}\left(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m\right)} \circ \times F_{Y_{(\overline{a}_m,0),k}|\overline{\ell}_m,\overline{a}_m, Y^{m+1:k-1}_{(\overline{a}_m,0)} = (y_{m+1},\cdots,y_{k-1})}(y_k) \ .$$

Many functions satisfy (8.1a). The restriction (8.1b) picks out a particular multivariate quantile-quantile function. Any alternative to (8.1b) that picked out a unique solution to (8.1a) would also serve for our purposes. Next recursively define, for $m = K-1, \ldots, 0$,

$$(8.2)\quad u_m\left(\overline{y}_{K+1}, \overline{\ell}_K, \overline{a}_K\right) = \gamma_m\left[\left\{y_{m+1}, u_{m+1}\left(\overline{y}_{K+1}, \overline{\ell}_K, \overline{a}_K\right)\right\}, \overline{\ell}_m, \overline{a}_m\right]$$

with $u_K\left(\overline{y}_{K+1}, \overline{\ell}_K, \overline{a}_K\right) \equiv \gamma_K\left(\underline{y}_{K+1}, \overline{\ell}_K, \overline{a}_K\right)$. Define

$$U_m = u_m\left(\overline{Y}_{K+1}, \overline{L}_K, \overline{A}_K\right).$$

Note U_m is a random vector of the same dimension as $\underline{Y}_{m+1}$.

The following theorem can be proved analogously to the proof of Theorem A1.1 in Robins (1993, Appendix 1).

THEOREM 8.1. $pr\left[U_m > \underline{y}_{m+1} \mid \overline{\ell}_m, \overline{a}_m\right] = pr\left[\underline{Y}_{(\overline{a}_{m-1},0),m+1} > \underline{y}_{m+1} \mid \overline{\ell}_m, \overline{a}_m\right]$.

We shall consider the model defined by the sole restriction that

$$(8.3a)\quad f\left[a_m \mid \overline{\ell}_m, \overline{a}_{m-1}, \underline{Y}_{(\overline{a}_{m-1},0),m+1} = \underline{y}_{m+1}\right] = \frac{t\left(a_m \mid \overline{\ell}_m, \overline{a}_{m-1}\right) \exp\left[q_m\left(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m\right)\right]}{\int t\left(a_m \mid \overline{\ell}_m, \overline{a}_{m-1}\right) \exp\left[q_m\left(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m\right)\right] d\mu\left(a_m\right)}$$

with

(8.3b) $q_m(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m)$ *known, satisfying* $q_m(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_{m-1}, a_m{=}0){=}0$

and

(8.3c) $t\left(a_m \mid \overline{\ell}_m, \overline{a}_{m-1}\right)$ *an unknown conditional density.*

The following theorem, proved in Appendix A, states that the SNDM model (8.3) is a NPI model.

THEOREM 8.2. *The model characterized by the sole restriction (8.3) on the law of* $\left(\overline{A}, \left\{\overline{L}_{\overline{a}}(K+1); \overline{a} \in \overline{\mathcal{A}}\right\}\right)$ *is a non-parametric model for the law of* $O = \left(\overline{L}_{K+1}, \overline{A}_K\right)$. *Furthermore, the functions* γ_m, *the density* $t(\cdot \mid \cdot, \cdot)$, *the law of* $\overline{Y}_{(0)}$, *and the conditional law* $\underline{Y}_{(\overline{A}_{m-1},0),m+1} \mid \overline{L}_m, \overline{A}_m$ *are all identified. Specifically, the identifying formulas are as follows. Let* γ_m^{-1} *be the inverse of the function* γ_m *with respect to the argument* $\underline{y}_{m+1}$. *Define* $U_m^* = (Y_m, U_m')'$; *note* U_m^* *is a vector of the dimension of* $\underline{Y}_m$. *Then by definition*

$$F_{U_{K+1}^* \mid \overline{L}_K, \overline{A}_K}\left(y_{K+1}\right) = F_{Y_{K+1} \mid \overline{L}_K, \overline{A}_K}\left(y_{K+1}\right).$$

We then have for $m = K, K-1, \ldots, 0$ *that* $\gamma_m^{-1}\left(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m\right)$ *is the unique solution to*

$$\text{(8.4a)}\quad F_{U^*_{m+1}|\bar{\ell}_m,\bar{a}_m}\left[\gamma_m^{-1}(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m)\right] = \tau(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m)/\tau(\infty, \bar{\ell}_m, \bar{a}_m)$$

satisfying

$$\text{(8.4b)}\qquad \gamma^{-1}_{m,k}\left(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m\right) \text{ depends on } \bar{y}_{K+1} \text{ only through } \bar{y}_k$$

where

$$\tau\left(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m\right) \equiv$$
$$\int_{-\infty}^{\underline{y}_{m+1}} f_{U^*_{m+1}|\bar{\ell}_m,\bar{a}_{m-1},a_m=0}\left(\underline{x}_{m+1}\right) \exp\left\{q_m\left(\underline{x}_{m+1}, \bar{\ell}_m, \bar{a}_m\right)\right\} d\underline{x}_{m+1} .$$

Furthermore,

$$\text{(8.5)}\qquad \begin{aligned} & f_{U_m|\bar{\ell}_m,\bar{a}_{m-1}}\left(\underline{y}_{m+1}\right) \\ &= f_{U^*_{m+1}|\bar{\ell}_m,\bar{a}_{m-1},a_m=0}\left(\underline{y}_{m+1}\right) \int_{-\infty}^{\infty} f\left[a_m \mid \bar{\ell}_m, \bar{a}_{m-1}\right] \\ &\times \exp\left[q_m(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m)\right]\left\{\tau\left(\infty, \bar{\ell}_m, \bar{a}_m\right)\right\}^{-1} d\mu(a_m) \end{aligned}$$

$$\text{(8.6)}\qquad t\left(a_m \mid \bar{\ell}_m, \bar{a}_{m-1}\right) = \frac{f\left(a_m \mid \bar{\ell}_m, \bar{a}_{m-1}\right) \tau\left(\infty, \bar{\ell}_m, \bar{a}_m\right)}{\int f\left(a_m \mid \bar{\ell}_m, \bar{a}_{m-1}\right) \tau\left(\infty, \bar{\ell}_m, \bar{a}_m\right) d\mu\left(a_m\right)}$$

$$f_{U^*_m|\bar{\ell}_{m-1},\bar{a}_{m-1}}\left(\underline{y}_{m+1}\right) =$$
$$\int f_{U_m|\bar{\ell}_m=(y_m,v_m,\bar{\ell}_{m-1}),\bar{a}_{m-1}}\left(\underline{y}_{m+1}\right) f\left(y_m, v_m \mid \bar{\ell}_{m-1}, \bar{a}_{m-1}\right) d\mu\left(v_m\right) .$$

REMARK 8.1. If we do not impose the equality restriction in (8.3b), Theorem (8.2) still holds provided we replace $q_m(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m)$ by $q_m(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m) - q_m(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_{m-1}, a_m = 0)$.

Theorem 8.2 has the following dual which is proved in Appendix A.

THEOREM 8.3. *The model for the joint law of* $\left(\overline{A}, \left\{\overline{L}_{\bar{a}}(K+1); \bar{a} \in \overline{\mathcal{A}}\right\}\right)$ *characterized by the sole restriction that*

$$\text{(8.7)}\qquad \gamma_m\left(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m\right) \text{ is known for each } m$$

with $\gamma_m(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m)$ *defined as in (8.1) is a non-parametric model for the distribution of the observed data* $O = \left(\overline{L}_{K+1}, \overline{A}_K\right)$. *Furthermore,*

$f\left[a_m \mid \bar{\ell}_m, \bar{a}_{m-1}, \underline{Y}_{(\bar{a}_{m-1,0}),m+1} = \underline{y}_{m+1}\right]$ *is identified and given by Eq. (8.3a) with*

$$\exp\left[q_m\left(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m\right)\right] = \frac{f_{U_m|\bar{\ell}_m,\bar{a}_m}\left[\underline{y}_{m+1}\right] f_{U_m|\bar{\ell}_m,\bar{a}_{m-1},a_m=0}\left(\underline{0}_{m+1}\right)}{f_{U_m|\bar{\ell}_m,\bar{a}_m}\left[\underline{0}_{m+1}\right] f_{U_m|\bar{\ell}_m,\bar{a}_{m-1},a_m=0}\left(\underline{y}_{m+1}\right)} . \tag{8.8}$$

In addition, $t\left(a_m \mid \bar{\ell}_m, \bar{a}_{m-1}\right)$ *and the densities of* $U_m \mid \bar{\ell}_m, \bar{a}_{m-1}$ *and* $U_m^* \mid \bar{\ell}_{m-1}, \bar{a}_{m-1}$ *are identified recursively by Eqs. (8.5) and (8.6).*

8.1.2. Structural nested mean models (SNMMs). For SNMMs, we shall redefine $\gamma_m\left(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m\right)$ and U_m. Robins (1994, 1997b) has previously considered two types of SNMMs: multiplicative SNMMs and additive SNMMs. For an additive SNMM, define

$$\gamma_m^*\left(\bar{\ell}_m, \bar{a}_m\right) = E\left[\underline{Y}_{(\bar{a}_m,0),m+1} - \underline{Y}_{(\bar{a}_{m-1},0),m+1} \mid \bar{\ell}_m, \bar{a}_m\right] \tag{8.9a}$$

and

$$\gamma_m\left(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m\right) = \underline{y}_{m+1} - \gamma_m^*\left(\bar{\ell}_m, \bar{a}_m\right) . \tag{8.9b}$$

For a multiplicative SNMM, define

$$\gamma_m\left(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m\right)' = \left(\gamma_{m,m+1}\left(y_{m+1}, \bar{\ell}_m, \bar{a}_m\right), \ldots, \gamma_{m,K+1}\left(y_{K+1}, \bar{\ell}_m, \bar{a}_m\right)\right) \tag{8.10a}$$

where

$$\gamma_{m,k}\left(y_k, \bar{\ell}_m, \bar{a}_m\right) = y_k \exp\left[-\left\{\gamma_{m,k}^*\left(\bar{\ell}_m, \bar{a}_m\right)\right\}\right] \tag{8.10b}$$

and

$$\gamma_{m,k}^*(\bar{\ell}_m, \bar{a}_m) = \log\left\{E[Y_{(\bar{a}_m,0),k} \mid \bar{\ell}_m, \bar{a}_m] / E[Y_{(\bar{a}_{m-1},0),k} \mid \bar{\ell}_m, \bar{a}_m]\right\} . \tag{8.10c}$$

Now define $u_m, U_m = (U_{m,m+1}, \ldots, U_{m,K+1})', U_m^*$ in terms of γ_m as above. Then we have for both additive and multiplicative SNMMs the following easily proved theorem.

THEOREM 8.4. *With* $\gamma_m\left(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m\right)$ *as defined in either (8.9) or (8.10),*

$$E\left[U_m \mid \bar{\ell}_m, \bar{a}_m\right] = E\left[\underline{Y}_{(\bar{a}_{m-1},0),m+1} \mid \bar{\ell}_m, \bar{a}_m\right] .$$

Next, consider the model defined by the sole restrictions that for $k = m+1, \ldots, K+1$ and $m = 0, \ldots, K$,

$$\text{(8.11a)} \quad E\left[Y_{(\bar{a}_{m-1},0),k} \mid \bar{\ell}_m, \bar{a}_m\right] = \Phi\left\{t\left(k, \bar{\ell}_m, \bar{a}_{m-1}\right) + q_m\left(k, \bar{\ell}_m, \bar{a}_m\right)\right\}$$

(8.11b) *with* $q_m(k, \bar{\ell}_m, \bar{a}_m)$ *known, satisfying* $q_m(k, \bar{\ell}_m, \bar{a}_{m-1}, a_m = 0) = 0$,

$$\text{(8.11c)} \qquad t\left(k, \bar{\ell}_m, \bar{a}_{m-1}\right) \text{ an unknown function}$$

and

$$\text{(8.11d)} \qquad \Phi(x) \text{ a known one-to-one function.}$$

Note $\Phi(x)$ no longer needs to be a distribution function. The following theorem, whose proof is omitted, gives conditions under which our model in a NPI model.

THEOREM 8.5. *Under Consistency Assumption 1, the model characterized by the sole restriction (8.11) is a non-parametric model for the law of* $O = \left(\overline{L}_{K+1}, \overline{A}_K\right)$
(a) with $\Phi(x)$ *the identity, provided* $E\left[Y_{k+1} \mid \bar{\ell}_k, \bar{a}_k\right]$ *may take values anywhere in* $(-\infty, \infty)$;
(b) with $\Phi(x) = \exp(x)$, *provided* $E\left[Y_{k+1} \mid \overline{L}_k, \overline{A}_k\right]$ *can take values anywhere in* $(0, \infty)$;
(c) with $\Phi(x) = e^x / \{1 + e^x\}$ *if the* Y_k *are dichotomous* $(0,1)$ *variables.*

Further, the functions γ^*_m (defined by either (8.9) or (8.10)) and t as well as $E\left[\overline{Y}_{(0)}\right]$ and the conditional expectations $E[\underline{Y}_{(\bar{a}_{m-1},0),m+1} \mid \bar{\ell}_m, \bar{a}_m]$ are identified. We give identifying formulae for two important cases. For $m = K, K-1, \ldots, 0$ and $k = m+1, \ldots, K+1$
(i) for an additive SNMM (i.e., γ^*_m given by (8.9)) with $\Phi(x) = x$

$$\begin{aligned}\text{(8.12a)} \quad \gamma^*_{m,k}\left(\bar{\ell}_m, \bar{a}_m\right) &= E\left[U^*_{m+1,k} \mid \bar{\ell}_m, \bar{a}_m\right] \\ &\quad - E\left[U^*_{m+1,k} \mid \bar{\ell}_m, \bar{a}_{m-1}, a_m = 0\right] - q_m\left(k, \bar{\ell}_m, \bar{a}_m\right)\end{aligned}$$

and $t\left(k, \bar{\ell}_m, \bar{a}_{m-1}\right) = E\left[U_{m,k} \mid \bar{\ell}_m, \bar{a}_{m-1}, a_m = 0\right]$,
(ii) for a multiplicative SNMM (i.e., γ^*_m given by (8.10)) with $\Phi(x) = e^x$

$$\begin{aligned}\text{(8.12b)} \quad \gamma^*_{m,k}\left(\bar{\ell}_m, \bar{a}_m\right) &= \log\left\{E\left[U^*_{m+1,k} \mid \bar{\ell}_m, \bar{a}_m\right]\right\} \\ &\quad - \log\left\{E\left[U^*_{m+1,k} \mid \bar{\ell}_m, \bar{a}_{m-1}, a_m = 0\right]\right\} - q_m\left(k, \bar{\ell}_m, \bar{a}_m\right)\end{aligned}$$

and

$$\text{(8.12c)} \qquad t\left(k, \bar{\ell}_m, \bar{a}_{m-1}\right) = \log\left\{E\left[U_{m,k} \mid \bar{\ell}_m, \bar{a}_{m-1}, a_m = 0\right]\right\}.$$

REMARK 8.2. Conditions (a)–(c) of Theorem 8.5 characterize a priori restrictions on the possible laws of O. For example, in the restriction (b),

we are considering all laws of the observed data O in which the Y_{k+1} have positive conditional means. We largely restrict attention to the additive SNMM with $\Phi(x) = x$ and the multiplicative SNMM with $\Phi(x) = e^x$ because these are the SNMMs for which we can later estimate the $\gamma_m^*(\bar{\ell}_m, \bar{a}_m)$ by the method of g-estimation based on modelling the law of A_m given $(\overline{A}_{m-1}, \overline{L}_m)$.

Theorem 8.5 has the following dual.

THEOREM 8.6. *If $E[Y_k \mid \overline{L}_k, \overline{A}_k]$ can take any value in $(-\infty, \infty)$, then the model characterized by the sole restriction that*

$$\gamma_m^*(\bar{\ell}_m, \bar{a}_m) \text{ is known for each } m \tag{8.13}$$

with $\gamma_m^*(\bar{\ell}_m, \bar{a}_m)$ as defined in (8.9) is a non-parametric model for the distribution F_O of the observed data. Furthermore, if $\Phi(x) = x$, then $q_m(k, \bar{\ell}_m, \bar{a}_m)$ defined by Eq. (8.11) is identified from Eq. (8.12a) applied recursively.

In addition, if $E[Y_k \mid \overline{L}_k, \overline{A}_k]$ can take any value in the interval $[0, \infty)$, then the model characterized by the sole restriction (8.13) with $\gamma_m^*(\bar{\ell}_m, \bar{a}_m)$ defined by (8.10) is a non-parametric model for the distribution of the observed data. Furthermore, if $\Phi(x) = e^x$, then $q_m(k, \bar{\ell}_m, \bar{a}_m)$ defined by (8.11a) is identified from Eq. (8.12b) applied recursively.

8.2. Inference and the curse of dimensionality.

8.2.1. Structural nested distribution models. In practice, due to the curse of dimensionality, in order to estimate, in a sensitivity analysis, $\gamma_m(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m)$ under model (8.3) when we vary the selection bias function $q_m(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m)$, we must consider a submodel of our NPI model (8.3) in which we impose parametric models

$$\gamma_m(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m) \in \left\{\gamma_m(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m; \psi); \psi \in \boldsymbol{\psi}\right\}, \tag{8.14}$$

$$t(a_m \mid \bar{\ell}_m, \bar{a}_{m-1}) \in \left\{t(a_m \mid \bar{\ell}_m, \bar{a}_{m-1}; \eta); \eta \in \boldsymbol{\eta}\right\} \tag{8.15}$$

where ψ and η are unknown finite dimensional parameters taking values in sets $\boldsymbol{\psi}$ and $\boldsymbol{\eta}$ and $\gamma_m(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m; \psi)$ and $t(a_m \mid \bar{\ell}_m, \bar{a}_{m-1}; \eta)$ are respectively a known function and a known density.

We then estimate (ψ, η) by g-estimation. That is, given a vector of functions $g_m(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m)$ of the dimension of ψ chose by the data analyst, we maximize an artificial likelihood (defined below) that depends on the selection bias function q_m of model (8.3), η, ψ and the artificial parameter θ. Specifically, we maximize

$$\prod_i \prod_m \mathcal{L}ik_{m,i}(\theta, \eta, \psi) \tag{8.16a}$$

with respect to η and θ with ψ held fixed to obtain estimates $\widehat{\eta}(\psi)$ and $\widehat{\theta}(\psi)$. We then define our g-estimate $\widehat{\psi}$ to be the value of ψ for which $\widehat{\theta}(\psi) = 0$ and define $\widehat{\eta} = \widehat{\eta}\left(\widehat{\psi}\right)$. Here

$$\mathcal{L}ik_{m,i}(\theta, \eta, \psi) = \nu_i(A_{m,i}) / \int \nu_i(a_m)\, d\mu(a_m) \tag{8.16b}$$

where $\nu(a_m) = t\left[a_m \mid \overline{L}_m, \overline{A}_{m-1}; \eta\right] \exp\left\{q_m\left[U_m(\psi), \overline{L}_m, \overline{A}_{m-1}, a_m\right] + \theta' g_m\left[U_m(\psi), \overline{L}_m, \overline{A}_{m-1}, a_m\right]\right\}$ and g_n is a user-supplied function, $U_m(\psi)$ is defined like U_m except with $\gamma_m\left(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m; \psi\right)$ in place of $\gamma_m\left(\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m\right)$. In conducting a sensitivity analysis, we will often choose a class of selection bias functions indexed by a parameter α of the form $q_m\left[\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m, \alpha\right]$ satisfying $q_m\left[\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m, 0\right] = 0$ so that $\alpha = 0$ corresponds to the absence of selection bias. We would then plot $\widehat{\psi}$ (or some functional of $\widehat{\psi}$) as a function of the selection bias parameter α. That is, we plot the function $\widehat{\psi}(\alpha)$ as the function of the selection bias parameter α. If our model characterized by (8.3), (8.14), and (8.15) is correctly specified for some particular α, then, for that α, $\widehat{\psi}(\alpha)$ will be a consistent asymptotically normal estimator of the true value of ψ. The optimal choice of g_m, say $g_{m,opt}$, is given in Sec. 9 of Robins (1997b) and results in a semiparametric efficient estimator of ψ under our model.

Consider the model in which we replace the assumption (8.3b) that the selection bias function is known with the weaker assumption that the true selection bias function is a member of the parametric family $\left\{q_m\left[\underline{y}_{m+1}, \overline{\ell}_m, \overline{a}_m, \alpha\right]; \alpha \in \boldsymbol{\alpha}\right\}$ where α is a finite dimensional parameter space. Since we have imposed restrictions (8.14), and (8.15), it is possible that (η, ψ, α) will be jointly identifiable. However, as discussed earlier, when the dimension of the parameters ψ and η are reasonably large, there will be little independent information about the three parameters, and their joint estimation will require truly huge sample sizes. Furthermore, the identification of α is strictly a consequence of our imposition of parametric models (8.14), and (8.15). Therefore, even with large sample sizes, if we carry out joint estimation of (α, ψ, η), any estimator of ψ will be highly sensitive to misspecification of both models (8.9) and (8.10), and there will be little power to detect such misspecification. As a consequence, we continue to recommend that one regard α as fixed and known and estimate (η, ψ) in a sensitivity analysis in which α is varied.

8.2.2. Structural nested mean models. For our SNMM, we consider a submodel of our NPI model (8.11) in which we impose parametric models for γ_m^* and for the conditional density of A_m given $\left(\overline{L}_m, \overline{A}_{m-1}\right)$. That is, we assume

$$\gamma_m^*\left(\overline{\ell}_m, \overline{a}_m\right) \in \left\{\gamma_m^*\left(\overline{\ell}_m, \overline{a}_m; \psi\right); \psi \in \boldsymbol{\psi}\right\} \tag{8.17}$$

and

$$f\left(a_m \mid \bar{\ell}_m, \bar{a}_{m-1}\right) \in \left\{f\left(a_m \mid \bar{\ell}_m, \bar{a}_{m-1}; \eta\right); \eta \in \boldsymbol{\eta}\right\} \tag{8.18}$$

with ψ and η unknown finite dimensional parameters. To estimate ψ by g-estimation under the model characterized by (8.11), (8.17), and (8.18), we first estimate η by the maximizer $\widehat{\eta}$ of the partial likelihood $\prod_{i=1}^{n} PL_i(\eta)$ where $PL(\eta) = \prod_{m=0}^{K} f\left[A_m \mid \overline{L}_m, \overline{A}_{m-1}; \eta\right]$. We then estimate ψ as follows. For $m = 0, \ldots, K$, let $g_m(a_m, \bar{\ell}_m, \bar{a}_{m-1})$ be a $(K+1-m) \times \dim \psi$ matrix-valued function chosen by the analyst and let $D'_m(\eta) = g'_m\left(A_m, \overline{L}_m, \overline{A}_{m-1}\right) - \int g'\left(a_m, \overline{L}_m, \overline{A}_{m-1}\right) dF\left(a_m \mid \bar{\ell}_m, \overline{A}_{m-1}; \eta\right)$. Then for $m = 0, \ldots, K$ and $k = m+1, \ldots, K+1$, let $c\left(k, \bar{\ell}_m, \bar{a}_{m-1}\right)$ be functions chosen by the analyst. For an additive SNMM with $\Phi(x) = x$, define $H_{m,k}(\psi) = U_{m,k}(\psi) - q_m\left(k, \overline{L}_m, \overline{A}_m\right) - c\left(k, \overline{L}_m, \overline{A}_{m-1}\right)$. Define $H_{m,k}(\psi) = U_{m,k}(\psi) \exp\left[-q_m\left(k, \overline{L}_m, \overline{A}_m\right) - c\left(k, \overline{L}_m, \overline{A}_{m-1}\right)\right]$ for a multiplicative SNMM with $\Phi(x) = e^x$. Here $q_m\left(k, \bar{\ell}_m, \bar{a}_m\right)$ is the known selection bias function in (8.11).

Then $\widehat{\psi}$ solves $0 = \sum_i W_i(\psi, \widehat{\eta})$ where $W(\psi, \eta) = \sum_{m=0}^{K} D'_m(\eta) \times \left(H_{m,m+1}(\psi), \ldots, H_{m,K+1}(\psi)\right)'$. It is easy to check that $E[W(\psi, \eta)] = 0$ at the true values of ψ and η under the model characterized by (8.11), (8.17), and (8.18). For readers conversant with the theory of semiparametric models, it will be of interest to note that in this model, the orthogonal complement to the nuisance tangent space is random vectors of the form $W(\psi, \eta) + S(\psi, \eta)$ where $S(\eta) = \sum_{m=0}^{K} r\left(\overline{A}_m, \overline{L}_m\right) - \int r\left(\overline{A}_m, \overline{L}_m\right) dF\left[A_m \mid \overline{L}_m, \overline{A}_{m-1}; \eta\right]$ for some user-supplied function $r(\cdot, \cdot)$.

It is interesting to note in the model characterized by (8.11), (8.17), and (8.18), in contrast to the model specified by (8.3), (8.14), and (8.15) studied in the last section, is that the density of A_m given $\overline{L}_m, \overline{A}_{m-1}$ is no longer ancillary when the selection bias function q_m is not identically zero. That is, the estimate of η obtained by maximizing the above partial likelihood is not an efficient estimator of η. The reason for this is that when the selection bias function q_m is not identically zero, $H_{m,k}(\psi)$ is not a deterministic function of $H_{m+1,k}(\psi), \overline{L}_{m+1}, \overline{A}_m$. In contrast, when the q_m are identically zero (i.e., there is no confounding), $H_{m,k}(\psi) = U_{m,k}(\psi)$ and $H_{m,k}(\psi)$ is therefore a deterministic function of the quantities mentioned above.

8.2.3. An alternative sensitivity analysis for continuous Y. The approach of Sec. 8.1.2 suggests the following alternative nonparametric identified model in the case of continuous Y. Define γ_m, $U_m = \left(U_{m,m+1}, \ldots, U_{m,K+1}\right)$ and U_m^* as in Sec. 8.1.1. Redefine $q_m\left(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m\right)$ as the unique solution to

$$F_{\underline{Y}_{(\overline{a}_{m-1},0),m+1}|\overline{\ell}_m,\overline{a}_{m-1},a_m=0}\left[q\left(\underline{y}_{m+1},\overline{\ell}_m,\overline{a}_m\right)\right] = F_{\underline{Y}_{(\overline{a}_{m-1},0),m+1}|\overline{\ell}_m,\overline{a}_m}\left(\underline{y}_{m+1}\right) \tag{8.19a}$$

satisfying

(8.19b) $q_{m,k}\left(\underline{y}_{m+1},\overline{\ell}_m,\overline{a}_m\right)$ *is a function of* $\overline{y}_{K+1}$ *only through* $\overline{y}_k$.

Let $H_m \equiv q_m\left(U_m,\overline{L}_m,\overline{A}_m\right)$. We then have the following obvious lemma.

LEMMA 8.1.

$$H_m \coprod A_m \mid \overline{L}_m,\overline{A}_{m-1} \tag{8.20}$$

and the following theorem.

THEOREM 8.7. *The model characterized by the sole restriction that, for each* m,

(8.21) $q_m\left(\underline{y}_{m+1},\overline{\ell}_m,\overline{a}_m\right)$ *defined by (8.19) is a known function*

is a non-parametric model for the law of F_O *of* O. *Furthermore, the functions* γ_m *are identified from the law of* F_O. *Specifically,* $\gamma_m = q_m^{-1} \circ \rho_m$, *where* $\rho_m\left(\underline{y}_{m+1},\overline{\ell}_m,\overline{a}_m\right)$ *is defined to be the function* $\gamma_m\left(\underline{y}_{m+1},\overline{\ell}_m,\overline{a}_m\right)$ *that would be obtained from the given law* F_O *under model (8.3) in the absence of confounding, i.e., when* $q_m\left(\underline{y}_{m+1},\overline{\ell}_m,\overline{a}_m\right)$ *as defined in (8.3b) is identically zero.*

REMARK 8.3. Since under the conditions of Theorem 8.7, γ_m is identified, it follows from Theorems 8.3 and 8.1 that under Assumption (8.21), the law of $\overline{Y}_{(0)}$ and the conditional laws of $\underline{Y}_{(\overline{A}_{m-1},0),m+1}$ given $\left(\overline{L}_m,\overline{A}_m\right)$ are identified.

Implications for *g*-estimation: It follows that given a parametric model $\gamma_m\left(\underline{y}_{m+1},\overline{\ell}_m,\overline{a}_m,\psi\right)$ for $\gamma_m\left(\underline{y}_{m+1},\overline{\ell}_m,\overline{a}_m\right)$ and a parametric model $f\left(a_m \mid \overline{\ell}_m,\overline{a}_{m-1};\eta\right)$ for $f\left(a_m \mid \overline{\ell}_m,\overline{a}_{m-1}\right)$, we can estimate ψ by standard g-estimation as in Robins (1997b). That is, $\widehat{\psi}$ is the vector ψ for which $\widehat{\theta}\left(\psi\right) = 0$ when maximizing (8.16a) with respect to θ and η with $\nu\left(a_m\right)$ in (8.16b) redefined to be

$$\nu(a_m) = f\left(a_m \mid \overline{L}_m,\overline{A}_{m-1};\eta\right)\exp\left\{\theta' g_m(H_m(\psi),\overline{L}_m,\overline{A}_{m-1},a_m)\right\} \tag{8.22}$$

with $H_m\left(\psi\right) = q_m\left(U_m\left(\psi\right),\overline{L}_m,\overline{A}_m\right)$. Again, if the function q_m is not identically zero (i.e., there is confounding), the partial likelihood estimator of η described following (8.18) is not efficient. This reflects the fact that $H_0\left(\psi\right)$ is not a deterministic function of $H_m\left(\psi\right),\overline{L}_m,\overline{A}_{m-1}$ except when the functions q_m are identically zero, in which case $H_m\left(\psi\right) = U_m\left(\psi\right)$. This also implies that the score equations for g-estimation described in Robins (1992) no longer have the uncorrelated increments property of a "martingale."

8.3. Identification of $\overline{Y}_{\overline{a}}$. Under the conditions of Theorem 8.2, 8.3, and 8.7, given a law F_O of the observed data, the distribution of $\overline{Y}_{\overline{a},K+1}$ is not identified except for $\overline{a} \equiv 0$. Similarly, in Theorems 8.5 and 8.6, the mean of $\overline{Y}_{\overline{a},K+1}$ is not identified except for $\overline{a} \equiv 0$. We shall now give sufficient and non-identifiable conditions to identify the above means and laws for any $\overline{a}$.

In fact, we give sufficient conditions to identify the mean and law of $\overline{Y}_{g,K+1}$, where $\overline{Y}_{g,K+1}$, as defined below, is the outcome under a possibly dynamic regime g.

Definition of Regimes: A treatment regime g is a collection of $K+1$ functions $g = (g_0, \ldots, g_K)$ where $g_m \equiv g_m\left(\overline{\ell}_m\right)$ maps an outcome history $\overline{\ell}_m$ into a treatment $g_m\left(\overline{\ell}_m\right) \in \mathcal{A}_m$. If, for each m, $g_m\left(\overline{\ell}_m\right)$ is a constant, say, a_m^* not depending on $\overline{\ell}_m$, we say the regime g is non-dynamic and write $g = \overline{a}^* = (a_0^*, \ldots, a_K^*)$. Otherwise we say the regime g is dynamic. The treatment at time m under a dynamic regime depends on the evolution of one's covariate history under that regime. We let $\mathcal{G}$ denote the set of all treatment regimes.

We let $\overline{L}_{g,K+1}$ be the subject's outcome history when, possibly contrary to fact, the subject follows regime g. Now define $g\left(\overline{\ell}_m\right) \equiv \left\{g_0\left(\ell_0\right), \ldots, g_m\left(\overline{\ell}_m\right)\right\}$ so $g\left(\overline{\ell}_m\right) \in \overline{\mathcal{A}}_m$. The counterfactual data $\overline{L}_{g,K+1}$ is linked to the observed data by the following consistency assumption.

Consistency Assumption 1:

$$\text{If } g\left(\overline{L}_m\right) = \overline{A}_m, \text{ then } \overline{L}_{g,m+1} = \overline{L}_{m+1} \,. \tag{8.23}$$

This consistency assumption states that if the subject has actually followed regime g until time t_{m+1}, then his counterfactual outcome under that regime and his observed outcome will agree through time t_{m+1}.

Consistency Assumption 2: If $g\left(\overline{\ell}_m\right) = g^*\left(\overline{\ell}_m\right)$ for regimes g and g^*, then $\overline{L}_{g,m+1} = \overline{L}_{g^*,m+1}$.

We shall now need to define various current treatment interaction functions. We adopt the convention

$$E\left[Z \mid \overline{\ell}_m, g\left(\overline{\ell}_m\right)\right] \equiv E\left[Z \mid \overline{L}_m = \overline{\ell}_m, \overline{A}_m = g\left(\overline{\ell}_m\right)\right]. \tag{8.24}$$

DEFINITION 8.1. *The additive current treatment interaction function* $r_m\left(\overline{\ell}_m, g\right)$ *is*

$$\begin{aligned} r_m\left(\overline{\ell}_m, g\right) &= E\left[\underline{Y}_{g,m+1} - \underline{Y}_{(g(\overline{\ell}_{m-1}),0),m+1} \mid \overline{\ell}_m, g\left(\overline{\ell}_m\right)\right] \\ &\quad - E\left[\underline{Y}_{g,m+1} - \underline{Y}_{(g(\overline{\ell}_{m-1}),0),m+1} \mid \overline{\ell}_m, g(\overline{\ell}_{m-1}), A_m \neq g_m(\overline{\ell}_m)\right]. \end{aligned} \tag{8.25}$$

If $r_m\left(\overline{\ell}_m, g\right) = 0$ for all $\overline{\ell}_m$, we say we have no additive current treatment interaction for regime g.

DEFINITION 8.2. *The multiplicative current treatment interaction function is* $r_m\left(\bar{\ell}_m, g\right) = \left(r_{m,m+1}\left(\bar{\ell}_m, g\right), \ldots, r_{m,K+1}\left(\bar{\ell}_m, g\right)\right)$ *where*

$$\begin{aligned} r_{m,k}\left(\bar{\ell}_m, g\right) = & \log\left\{E\left(Y_{g,k} \mid \bar{\ell}_m, g\left(\bar{\ell}_m\right)\right)\right\} \\ & - \log\left\{E\left(Y_{\left(g\left(\bar{\ell}_{m-1}\right),0),k} \mid \bar{\ell}_m, g\left(\bar{\ell}_m\right)\right)\right\} \\ & - \Big[\log\left\{E\left(Y_{g,k} \mid \bar{\ell}_m, g\left(\bar{\ell}_{m-1}\right), A_m \neq g_m\left(\bar{\ell}_m\right)\right)\right\} \\ & - \log\left\{E\left(Y_{\left(g\left(\bar{\ell}_{m-1}\right),0),k} \mid \bar{\ell}_m, g\left(\bar{\ell}_{m-1}\right), A_m \neq g_m\left(\bar{\ell}_m\right)\right)\right\}\Big] . \end{aligned}$$

DEFINITION 8.3. *Define* $\nu_m^{\dagger}\left(\underline{y}_{m+1}, \bar{\ell}_m, g\right)$ *to be the unique solution to*

$$\begin{aligned} & F_{\underline{Y}_{\left(g\left(\bar{\ell}_{m-1}\right),0),m+1} \mid \bar{\ell}_m, g\left(\bar{\ell}_{m-1}\right), A_m \neq g_m\left(\bar{\ell}_m\right)}\left[\nu_m^{\dagger}\left(\underline{y}_{m+1}, \bar{\ell}_m, \bar{a}_m\right)\right] \\ & \quad = F_{\underline{Y}_{g,m+1} \mid \bar{\ell}_m, g\left(\bar{\ell}_{m-1}\right), A_m \neq g_m\left(\bar{\ell}_m\right)}\left(\underline{y}_{m+1}\right) \end{aligned} \tag{8.26a}$$

satisfying

$$\nu_{m,k}^{\dagger}\left(\underline{y}_{m+1}, \bar{\ell}_m, g\right) \text{ is a function of } \bar{y}_{K+1} \text{ only through } \bar{y}_k \; . \tag{8.26b}$$

Define $\nu_m\left(\underline{y}_{m+1}, \bar{\ell}_m, g\right)$ similarly, except with conditioning events being $\bar{\ell}_m, g\left(\bar{\ell}_m\right)$.

DEFINITION 8.4. *The distribution current treatment interaction function is* $r_m\left(\underline{y}_{m+1}, \bar{\ell}_m, g\right) \equiv r_m \equiv \nu_m^{\dagger} \circ \nu_m^{-1}$ *where* ν_m^{-1} *is the inverse of the function* ν_m *with respect to the argument* $\underline{y}_{m+1}$. *We now state our main theorems.*

THEOREM 8.8. *Under the consistency assumption* 1, *given (i) a law* F_O *of the observed data* $O = \left(\bar{A}_K, \bar{L}_{K+1}\right)$, *(ii) the conditional means* $E\left[\underline{Y}_{\left(g\left(\bar{\ell}_{m-1}\right),0),m+1} \mid \bar{\ell}_m, \bar{a}_m\right]$ *and (iii) either the additive or multiplicative current treatment interaction functions* $r_m\left(\bar{\ell}_m, g\right)$, *then* $E\left[\overline{Y}_{g,K+1}\right]$ *and* $E\left[\underline{Y}_{g,m+1} \mid \bar{\ell}_m, g\left(\bar{\ell}_m\right)\right]$ *are determined (i.e., identified).*

THEOREM 8.9. *Under the consistency assumption* 1, *given (i) a law* F_O *of the observed data* O, *(ii) the conditional laws* $\underline{Y}_{\left(g\left(\bar{\ell}_{m-1}\right),0),m+1}$ *given* $\left(\bar{\ell}_m, \bar{a}_m\right)$, *and the distribution current treatment interaction functions* $r_m\left(\underline{y}_{m+1}, \bar{\ell}_m, g\right)$, *the law of* $\overline{Y}_{g,K+1}$, *and the conditional laws of* $\underline{Y}_{g,m+1}$ *given* $\left(\bar{\ell}_m, g\left(\bar{\ell}_m\right)\right)$ *are determined (i.e., identified).*

Proof of Theorem 8.8. We first note that

$$E\left[Y_{g,m+1} \mid \bar{\ell}_m, g\left(\bar{\ell}_m\right)\right] = E\left[Y_{m+1} \mid \bar{\ell}_m, g\left(\bar{\ell}_m\right)\right]$$

by Consistency Assumption 1. Further, $E\left[Y_m \mid \bar{\ell}_m, g\left(\bar{\ell}_m\right)\right]$ is identified, since F_O is given. Thus, arguing recursively in reverse, beginning at $m = K$, if we can show that

$$E\left[\underline{Y}_{g,m+1} \mid \bar{\ell}_m, g\left(\bar{\ell}_m\right)\right] \text{ identified} \tag{8.27}$$

implies, under our assumptions, that $E\left[\underline{Y}_{g,m+1} \mid \bar{\ell}_m, g\left(\bar{\ell}_{m-1}\right)\right]$ is identified, then the theorem is proved since *(i)*, by F_O given, we can then conclude that $E\left[\underline{Y}_{g,m+1} \mid \bar{\ell}_{m-1}, g\left(\bar{\ell}_{m-1}\right)\right]$ is identified, and thus, by the first display in the proof, $E\left[\underline{Y}_{g,m} \mid \bar{\ell}_{m-1}, g\left(\bar{\ell}_{m-1}\right)\right]$ is identified. Now,

$$\begin{aligned} &E\left[\underline{Y}_{g,m+1} \mid \bar{\ell}_m, g(\bar{\ell}_{m-1})\right] \\ &= E\left[\underline{Y}_{g,m+1} \mid \bar{\ell}_m, g(\bar{\ell}_m)\right] pr\left[A_m = g(\bar{\ell}_m) \mid \bar{\ell}_m, g(\bar{\ell}_{m-1})\right] \\ &+ E\left[\underline{Y}_{g,m+1} \mid \bar{\ell}_m, g(\bar{\ell}_{m-1}), A_m \neq g_m(\bar{\ell}_m)\right] pr\left[A_m \neq g_m(\bar{\ell}_m) \mid \bar{\ell}_m, g(\bar{\ell}_{m-1})\right] \end{aligned} \tag{8.28}$$

Given (8.27) and F_O, $E\left[\underline{Y}_{g,m+1} \mid \bar{\ell}_m, g\left(\bar{\ell}_{m-1}\right), A_m \neq g_m\left(\bar{\ell}_m\right)\right]$ is the only unknown on the RHS of Eq. (8.27). However, $E\left[\underline{Y}_{g,m+1} \mid \bar{\ell}_m, g\left(\bar{\ell}_{m-1}\right), A_m \neq g_m\left(\bar{\ell}_m\right)\right]$ is a function of $E\left[\underline{Y}_{g,m+1} \mid \bar{\ell}_m, g\left(\bar{\ell}_m\right)\right]$, the given F_O, $r_m\left(\bar{\ell}_m, g\right)$ and $E\left[Y_{\left(g\left(\bar{\ell}_{m-1}\right),0\right)} \mid \bar{\ell}_m, \bar{a}_m\right]$ and thus is determined (i.e., identified), which completes the proof. □

The proof of Theorem 8.9 is similar and is omitted.

We next take up the question whether specifying particular current treatment interaction functions places any restrictions on the joint law of the observed data.

DEFINITION 8.5. *We say that any given function $r_m\left(\bar{\ell}_m, g\right)$ is a potential additive or multiplicative current treatment interaction function if $r_m\left(\bar{\ell}_m, g\right)$ satisfies*

$$\text{if } g_k\left(\bar{\ell}_k\right) = 0 \text{ for } k \geq m \text{, then } r_m\left(\bar{\ell}_m, g\right) = 0. \tag{8.29a}$$

$$g_m^*\left(\bar{\ell}_m\right) = g_m\left(\bar{\ell}_m\right) \text{ then } r_{m,m+1}\left(\bar{\ell}_m, g\right) = r_{m,m+1}\left(\bar{\ell}_m, g^*\right). \tag{8.29b}$$

Note that the Equations (8.29a) and (8.29b) hold for any additive or multiplicative current treatment interaction function by the Consistency Assumptions 1 and 2.

REMARK 8.4. Any function satisfying (8.29) can be represented as follows. Given a collection of functions $g = (g_0, \ldots, g_K)$ as defined previously, let $r_m^*\left(\bar{\ell}_m, \bar{a}_m, g\right) = \left(r_{m,m+1}^*\left(\bar{\ell}_m, \bar{a}_m\right), r_{m,m+2}^*\left(\bar{\ell}_m, \bar{a}_m, g_{m+1}\right), \ldots, r_{m,K+1}^*\left(\bar{\ell}_m, \bar{a}_m, g_{m+1}, \ldots, g_K\right)\right)$. Then the set of functions $r_m\left(\bar{\ell}_m, g\right)$ satisfying (8.29) is precisely the set $\{r_m^*\left(\bar{\ell}_m, g\left(\bar{\ell}_m\right), g\right); r_{m,k}^*\left(\bar{\ell}_m, \bar{a}_m, g_{m+1}, \ldots, g_k\right) = 0$ *if* $a_m = 0$ and $g_{m+1}, \ldots, g_k$ *are all the zero function*$\}$.

Similarly,

DEFINITION 8.6. *We say a function $r_m\left(\underline{y}_{m+1}, \bar{\ell}_m, g\right)$ is a potential current treatment interaction function if (i) $r_{m,k}\left(\underline{y}_{m+1}, \bar{\ell}_m, g\right)$ depends on*

$\overline{y}_{K+1}$ only through $\overline{y}_k$ and is increasing in y_k, (ii) $r_m(\underline{y}_{m+1}, \overline{\ell}_m, g) = \underline{y}_{m+1}$ if $g_k(\overline{\ell}_k) = 0$ for $k \geq m$, and (iii) if $g^*(\overline{\ell}_m) = g(\overline{\ell}_m)$ then $r_{m,m+1}(\underline{y}_{m+1}, \overline{\ell}_m, g) = r_{m,m+1}(\underline{y}_{m+1}, \overline{\ell}_m, g^*)$.

Consider the following theorem.

THEOREM 8.10. *Under Consistency Assumptions 1 and 2, given any law F_O for the observed data, any function $r_m(\overline{\ell}_m, g)$ satisfying (8.29), and any function $q_m(k, \overline{\ell}_m, \overline{a}_m)$, there exists a joint law for $(\overline{A}, \{\overline{L}_{g,K+1};\ g \in \mathcal{G}\})$ satisfying model (8.11) with $\Phi(x) = x$ and with $r_m(\overline{\ell}_m, g)$ the additive current treatment interaction function, provided $E[Y_{k+1} \mid \overline{\ell}_k, \overline{a}_k]$ may take values anywhere in $(-\infty, \infty)$.*

REMARK 8.5. Note by Consistency Assumption 2, $\{\overline{L}_{\overline{a},K+1}; \overline{a} \in \overline{\mathcal{A}}\}$ completely determines $\overline{L}_{g,K+1}$ for each $g \in \mathcal{G}$. Thus, in the statement of Theorem 8.10, we could have replaced $\{\overline{L}_{g,K+1}; g \in \mathcal{G}\}$ by $\{\overline{L}_{\overline{a},K+1}; \overline{a} \in \overline{\mathcal{A}}\}$.

REMARK 8.6. Note, it follows from Theorem 8.5 that under the conditions of Theorem 8.10, $\gamma_m^*(\overline{\ell}_m, \overline{a}_m)$ defined by (8.9a) as identified as a function of $q_m(k, \overline{\ell}_m, \overline{a}_m)$. Suppose for a particular choice of q_m, $\gamma_m^*(\overline{\ell}_m, \overline{a}_m)$ is identically zero. This implies according to (8.9a) that there is no effect of a final brief blip of treatment a_m at time t_m on subjects with history $\overline{\ell}_m, \overline{a}_m$. Often it would seem reasonable to have strong prior beliefs $\gamma_m^*(\overline{\ell}_m, \overline{a}_m) = 0$, then the g-null hypothesis that $E[Y_{g,K+1}]$ is the same for all $g \in \mathcal{G}$ was true. This prior will be satisfied if we choose the potential additive current treatment interaction function $r_m(\overline{\ell}_m, g)$ in Theorem 8.10 to be identically zero whenever $\gamma_m^*(\overline{\ell}_m, \overline{a}_m)$ is identically zero.

For multiplicative current treatment interaction functions, we have the following similar weaker result.

THEOREM 8.11. *Under Consistency Assumptions 1 and 2, given a law F_O, a function $q_m(k, \overline{\ell}_m, \overline{a}_m)$, and a function $r_m(\overline{\ell}_m, g)$ satisfying Eq. (8.29), there exists a joint law for $(\overline{A}, \{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\})$ satisfying (8.11) with $\Phi(x) = e^x$ and with $r_m(\overline{\ell}_m, g)$ the multiplicative current treatment interaction function for non-dynamic regimes $g = \overline{a}$, provided $E[Y_{k+1} \mid \overline{\ell}_k, \overline{a}_k]$ may take values anywhere in $[0, \infty)$.*

REMARK 8.7. Unlike Theorem 8.10, Theorem 8.11 cannot be extended to include dynamic regimes. Specifically, if $E[Y_{g,k} \mid \overline{\ell}_m, g(\overline{\ell}_m)]$ only takes values in $[0, \infty)$, then $E[\underline{Y}_{g,m+2} \mid \overline{\ell}_m, g(\overline{\ell}_{m-1}), A_m \neq g_m(\overline{\ell}_m)]$ must be zero if

$$E\left[\underline{Y}_{(g(\overline{\ell}_m),a_{m+1},0),m+2} \mid \overline{\ell}_m, g(\overline{\ell}_{m-1}), A_m \neq g(\overline{\ell}_m)\right] = 0$$
$$\text{for all } a_{m+1} \in \mathcal{A}_{m+1}\ .$$

This follows by the fact that, under Consistency Assumption 2, the last display implies $\underline{Y}_{g,m+2} = 0$ on $L_m, \overline{A}_{m-1} = g(\overline{\ell}_{m-1}), A_m \neq g(\overline{\ell}_m)$.

It follows that knowledge of $r_m(\bar{\ell}_m, g)$ for non-dynamic regimes can place various kinds of restrictions on $r_m(\bar{\ell}_m, g)$ for dynamic regimes.

For similar reasons, for continuous Y, the best we can obtain is the following for continuous $\overline{Y}$.

THEOREM 8.12. *Under Consistency Assumptions 1 and 2, given a law F_O and a potential current treatment interaction function $r_m(\underline{y}_{m+1}, \bar{\ell}_m, g)$, there exists a joint law for $(\overline{A}, \{\overline{L}_{(\bar{a})}; \bar{a} \in \overline{\mathcal{A}}\})$ satisfying the constraints imposed by either model (8.3) or model (8.21) with $r_m(\underline{y}_{m+1}, \bar{\ell}_m, g)$ the current treatment interaction function for non-dynamic regimes $g = \bar{a}$.*

REMARK 8.8. If $K = 0$ so we only have a single time-independent treatment A_0, Theorems 8.11 and 8.12 are true with $r_m(\bar{\ell}_m, g)$ and $r_m(\underline{y}_{m+1}, \bar{\ell}_m, g)$ the multiplicative and distribution current treatment interaction functions for all regimes g, both dynamic and non-dynamic. Here of course m takes only the value 0.

8.4. Logistic structural nested mean models. In this subsection, we introduce Logistic Structural Nested Mean Models (SNMMs) for dichotomous Y_k. Redefine

$$\gamma^*_{m,k}(\bar{\ell}_m, \bar{a}_m) = \begin{array}{l} logit\{E[Y_{(\bar{a}_m,0),k} \mid \bar{\ell}_m, \bar{a}_m]\} \\ - logit\{E[Y_{(\bar{a}_{m-1},0)} \mid \bar{\ell}_m, \bar{a}_m]\} \ . \end{array} \tag{8.30}$$

Then Theorem 8.5 is obviously still true with γ^*_m redefined by (8.30). Indeed, the identifying formulas (8.12b) and (8.12c) remain true with "*log*" replaced by "*logit*" and $U^*_{m+1,k}$ and $U_{m,k}$ replaced by $Y_{(\bar{a}_m,0),k}$ and $Y_{(\bar{a}_{m-1},0),k}$ respectively. Note that we obtain identification because $Y_{\bar{a}_K,K+1} = Y_{K+1}$ by Consistency Assumption 1.

However, there is no longer a function U_m of $\underline{Y}_{m+1}$ and the redefined γ^*_m such that Theorem 8.4 holds. Thus, we cannot estimate a parametric logistic SNMM $\gamma^*_m(\bar{\ell}_m, \bar{a}_m, \psi)$ by g-estimation. See Remark 8.13 in Sec. 8.5 for further discussion of this point. Fully parametric likelihood based and Bayesian approaches to estimation of this model are considered in Secs. 8.5 and 11.

Now define the logistic current treatment interaction function like a multiplicative current treatment interaction function except with "*logit*" replacing "*log*" in the definition. Then Theorem 8.8 remains true with $r_m(\bar{\ell}_m, g)$ the logistic current treatment interaction function. Further, the following analog of Theorem 8.11 holds.

THEOREM 8.13. *Suppose Y_k is dichotomous. Under Consistency Assumptions 1 and 2, given a law F_O, a function $q_m(k, \bar{\ell}_m, \bar{a}_m)$, and a function $r_m(\bar{\ell}_m, g)$ satisfying Eq. (8.29), there exists a joint law for $(\overline{A}, \{\overline{L}_{\bar{a}}; \bar{a} \in \overline{\mathcal{A}}\})$ satisfying (8.11) with $\Phi(x) = e^x/(1+e^x)$ and with $r_m(\bar{\ell}_m, g)$ the logistic current treatment interaction function for non-dynamic regimes $g = \bar{a}$.*

8.5. Likelihood inference for structural nested models. The likelihood function for the structural nested distribution model of Sec. 8.1.1 has been given in Robins (1997b) and is quite straightforward. We shall next, therefore, consider the likelihood function for structural nested mean models. In the following, $\Phi^{-1}(x)$ is x, $\log x$, and *logit* $x = \log\{x/(1-x)\}$ for additive multiplicative and logistic models respectively. Thus, $\Phi(x)$ is respectively x, e^x, and $e^x/(1+e^x)$. Recall the definitions:

$$\begin{aligned}\gamma^*_{m,k+1}(\bar{\ell}_m, \bar{a}_m) = {} & \Phi^{-1}\{E[Y_{(\bar{a}_m,0),k+1} \mid \bar{\ell}_m, \bar{a}_m]\} \\ & - \Phi^{-1}\{E[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_m, \bar{a}_m]\}\end{aligned} \tag{8.31}$$

$$\begin{aligned}q_m(k+1, \bar{\ell}_m, \bar{a}_m) = {} & \Phi^{-1}\{E[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_m, \bar{a}_m]\} \\ & - \Phi^{-1}\{E[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_m, \bar{a}_{m-1}, a_m = 0]\}.\end{aligned} \tag{8.32}$$

Define

$$\begin{aligned}\nu^*_m(k+1, \bar{\ell}_m, \bar{a}_{m-1}) = {} & \Phi^{-1}\{E[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_m, \bar{a}_{m-1}]\} \\ & - \Phi^{-1}\{E[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_{m-1}, \bar{a}_{m-1}, \ell_m = 0]\}.\end{aligned} \tag{8.33}$$

To construct the likelihood function, we need an expression for the following.

$$\begin{aligned}&\Phi^{-1}\{E[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_m, \bar{a}_m]\} - \Phi^{-1}\{E[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_m, \bar{a}_{m-1}]\} \\ &\qquad \equiv q_m(k+1, \bar{\ell}_m, \bar{a}_m) - \Gamma(k+1, \bar{\ell}_m, \bar{a}_{m-1})\end{aligned} \tag{8.34}$$

where

$$\begin{aligned}\Gamma(k+1, \bar{\ell}_m, \bar{a}_{m-1}) \equiv {} & -\Phi^{-1}\{E(Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_m, \bar{a}_{m-1}, a_m = 0)\} \\ & + \Phi^{-1}\{E[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_m, \bar{a}_{m-1}]\},\end{aligned} \tag{8.35}$$

and

$$\begin{aligned}&\Phi^{-1}\{E[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_m, \bar{a}_{m-1}]\} - \Phi^{-1}\{E[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_{m-1}, \bar{a}_{m-1}]\} \\ &\qquad \equiv \nu^*_m(k+1, \bar{\ell}_m, \bar{a}_{m-1}) - \Gamma^*(k+1, \bar{\ell}_{m-1}, \bar{a}_{m-1})\end{aligned} \tag{8.36}$$

where

$$\begin{aligned}\Gamma^*(k+1, \bar{\ell}_{m-1}, \bar{a}_{m-1}) = {} & -\Phi^{-1}\{E[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_{m-1}, \ell_m = 0, \bar{a}_{m-1}]\} \\ & + \Phi^{-1}\{E[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_{m-1}, \bar{a}_{m-1}]\}.\end{aligned} \tag{8.37}$$

Note for $\Phi(x) = e^x$ *or* $\Phi(x) = x$, one can calculate that

$$\Gamma\left(k+1, \bar{\ell}_m, \bar{a}_{m-1}\right) = \Phi^{-1} \int \Phi\left\{q_m\left(k+1, \bar{\ell}_m, \bar{a}_m\right)\right\} dF\left(a_m \mid \bar{\ell}_m, \bar{a}_{m-1}\right) \tag{8.38}$$

and

$$\Gamma^*\left(k+1, \bar{\ell}_{m-1}, \bar{a}_{m-1}\right) = \Phi^{-1} \int \Phi\left[\nu_m^*\left(k+1, \bar{\ell}_m, \bar{a}_{m-1}\right)\right] dF\left[\ell_m \mid \bar{\ell}_{m-1}, \bar{a}_{m-1}\right] . \tag{8.39}$$

For $\Phi(x) = e^x/(1+e^x)$, $\exp\left\{-\Gamma^*\left(k+1, \bar{\ell}_{m-1}, \bar{a}_{m-1}\right)\right\}$ is the unique solution x to the equation

$$(1-p)^2 = \int \left\{(1-p)\nu_m^*(k+1, \bar{\ell}_m, \bar{a}_{m-1})^{-1} x^{-1} + p\right\}^{-1} dF\left[\ell_m \mid \bar{\ell}_{m-1}, \bar{a}_{m-1}\right] \tag{8.40}$$

with

$$p \equiv E\left[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_{m-1}, \bar{a}_{m-1}\right] . \tag{8.41}$$

Further, $\exp\left[-\Gamma\left(k+1, \bar{\ell}_m, \bar{a}_{m-1}\right)\right]$ is the unique solution x to

$$(1-p)^2 = \int \left\{(1-p)\, q_m\left(k+1, \bar{\ell}_m, \bar{a}_m\right)^{-1} x^{-1} + p\right\}^{-1} dF\left[a_m \mid \bar{\ell}_m, \bar{a}_{m-1}\right] \tag{8.42}$$

with

$$p \equiv E\left[Y_{(\bar{a}_{m-1},0),k+1} \mid \bar{\ell}_m, \bar{a}_{m-1}\right] . \tag{8.43}$$

To construct the likelihood function for non-parametric, semiparametric, and/or parametric versions of our model, we shall consider sets of densities indexed by (possibly infinite dimensional parameters) η and $\gamma' = (\gamma_1', \gamma_2')$

$$\left\{f\left[a_m \mid \bar{\ell}_m, \bar{a}_{m-1}; \eta\right];\ \eta \in \boldsymbol{\eta},\ 0 \le m \le K\right\} \tag{8.44a}$$

$$\left\{f\left(\varepsilon_{k+1} \mid \bar{\ell}_k, \bar{a}_k; \gamma_1\right);\ \int \varepsilon_{k+1} dF\left(\varepsilon_{k+1} \mid \bar{\ell}_k, \bar{a}_k; \gamma_1\right) = 0 \text{ and } \gamma_1 \in \boldsymbol{\gamma}_1,\ -1 \le k \le K\right\} \tag{8.44b}$$

$$\left\{f\left(v_{k+1} \mid \bar{\ell}_k, \bar{a}_k, Y_{k+1}; \gamma_2\right);\ \gamma_2 \in \boldsymbol{\gamma}_2,\ -1 \le k \le K\right\} . \tag{8.44c}$$

We use the convention that for any $\overline{Z}$, $\overline{Z}_{-1} \equiv 0$ with probability 1. We shall also need to consider sets of functions indexed by parameters ψ, μ_1, and μ_2.

$$\left\{\gamma^*_{m,k+1}\left(\overline{\ell}_m, \overline{a}_m; \psi\right);\ \gamma^*_{m,k+1}\left(\overline{\ell}_m, \overline{a}_{m-1}, a_m = 0; \psi\right) = 0, \quad \psi \in \boldsymbol{\psi},\ K \geq k \geq m,\ 0 \leq m \leq K\right\} \tag{8.44d}$$

$$\left\{q_m(k+1, \overline{\ell}_m, \overline{a}_m; \mu_1);\ q_m(k+1, \overline{\ell}_m, \overline{a}_{m-1}, a_m = 0; \mu_1) = 0, \quad \mu_1 \in \boldsymbol{\mu}_1,\ K \geq k \geq m,\ 0 \leq m \leq K\right\} \tag{8.44e}$$

$$\left\{\nu^*_m(k+1, \overline{\ell}_1, \overline{a}_{m-1}; \mu_2);\ \nu^*_m(k+1, \overline{\ell}_{m-1}, \ell_m = 0, \overline{a}_{m-1}; \mu_2) = 0, \quad \mu_2 \in \boldsymbol{\mu}_2,\ K \geq k \geq m,\ 0 \leq m \leq K\right\}. \tag{8.44f}$$

Finally, we need to consider a set of vectors

$$\left\{\omega = (\omega_0, \ldots, \omega_{K+1});\, \omega \in \boldsymbol{\omega} \subset R^{K+2}\right\}. \tag{8.44g}$$

We shall consider a model in which the unknown parameters, functions, and densities lie in the sets specified in (8.44). We derive the likelihood function for our model by considering the following algebraic identity, $K \geq k \geq 0$.

$$\begin{aligned}\Phi^{-1}\left\{E\left(Y_{k+1} \mid \overline{\ell}_k, \overline{a}_k\right)\right\} &= \sum_{m=0}^{k} \gamma^*_{m,k+1}\left(\overline{\ell}_m, \overline{a}_m\right) \\ &+ \left\{q_m\left(k+1, \overline{\ell}_m, \overline{a}_m\right) - \Gamma\left(k+1, \overline{\ell}_m, \overline{a}_{m-1}\right)\right\} \\ &+ \left\{\nu^*_m\left(k+1, \overline{\ell}_m, \overline{a}_{m-1}\right) - \Gamma^*\left(k+1, \overline{\ell}_{m-1}, \overline{a}_{m-1}\right)\right\} \\ &+ E\left(Y_{(0),k+1}\right).\end{aligned} \tag{8.45}$$

Write, for $k = -1, \ldots, K$,

$$\varepsilon_{k+1} = Y_{k+1} - E\left(Y_{k+1} \mid \overline{L}_k, \overline{A}_k\right). \tag{8.46}$$

Then the likelihood for $O = \left(\overline{A}_K, \overline{L}_{K+1}\right)$ with $L_k \equiv (V_k, Y_k)$ can be written as follows, in terms of the parameter $\rho = (\psi, \eta, \gamma, \mu, \omega)$.

$$\begin{aligned}f[O; \rho] = &\prod_{m=0}^{K+1} f\left[\varepsilon_k(\rho) \mid \overline{L}_{k-1}, \overline{A}_{k-1}; \gamma_1\right] f\left(V_k \mid \overline{L}_{k-1}, \overline{A}_{k-1}, Y_k; \gamma_2\right) \\ &\times \prod_{m=0}^{K} f\left[A_k \mid \overline{L}_k, \overline{A}_{k-1}; \eta\right]\end{aligned} \tag{8.47}$$

where, for $k = -1, \ldots, K$,

$$\varepsilon_{k+1}(\rho) = Y_{k+1} - E\left(Y_{k+1} \mid \overline{L}_k, \overline{A}_k; \rho\right), \tag{8.48}$$

$\Phi^{-1}\left\{E\left[Y_0 \mid \overline{L}_{-1}, \overline{A}_{-1}; \rho\right]\right\} = \omega_0$ and $\Phi^{-1}\left\{E\left[Y_{k+1} \mid \overline{L}_k, \overline{A}_k; \rho\right]\right\}$ for $k \geq 0$ is given by the RHS of (8.45) except with the parameterized versions of γ^*, q, and ν^*, as defined in (8.44), replacing the unparameterized versions and ω_{k+1} replacing $\Phi^{-1}\left\{E\left[Y_{(0),k+1}\right]\right\}$.

Note by (8.38), (8.39), (8.40), and (8.42), Γ and Γ^* on the RHS of (8.45) can be parameterized in terms of the parameter ρ, although, in the logistic case, the dependence on ρ will be quite complex. Note that if Y_k is dichotomous and $\Phi(x) = e^x/(1+e^x)$, there is no parameter γ_1, since the conditional law of $\varepsilon_{k+1} \mid \overline{\ell}_k, \overline{a}_k$ is determined by the remaining components of ρ.

REMARK 8.9. The parameter ρ will have variation-independent components (i.e., ρ will take values in the product space $\boldsymbol{\rho} = \boldsymbol{\psi} \times \boldsymbol{\eta} \times \boldsymbol{\gamma} \times \boldsymbol{\mu} \times \boldsymbol{\omega}$) if *(i)* $\Phi(x) = e^x/(1+e^x)$ and Y_k is dichotomous or *(ii)* $\Phi(x) = x$ and Y_k is absolutely continuous with respect to Lesbesgue measure with support on the entire real line. If $\Phi(x) = e^x$ and Y_k is absolutely continuous with respect to Lesbesgue measure with support on the positive half line, then $\boldsymbol{\rho}$ will be variation independent if we (a) redefine $\varepsilon_{k+1} = Y_{k+1}/E\left(Y_{k+1} \mid \overline{L}_k, \overline{A}_k\right)$, (b) restrict the integral in (8.44b) to be 1 rather than 0 and require ε_{k+1} to be supported on the positive half line, and (c) add the Jacobian terms $\partial\, \varepsilon_k/\partial Y_k$ to the likelihood (8.47).

Now consider the model in which the parameter space for $\rho = (\psi, \eta, \gamma, \mu, \omega)$ is unrestricted in the sense that the spaces $\boldsymbol{\psi}, \boldsymbol{\eta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\omega}$ are as large as possible, subject to the restrictions specified in (8.44). In this model, ψ, μ_1, μ_2, and ω are not identified.

REMARK 8.10. Consider the model in which q_m is known [i.e., the set $\boldsymbol{\mu}_1$ in (8.44) has a single element] but the other components of ρ remain unrestricted as above. It then follows from Theorem 8.5 and Sec. 8.4 that the remainder of ρ is identified, since this is a NPI model for the law F_O of the observed data if *(a)* $\Phi(x) = x$ and $E\left(Y_{k+1} \mid \overline{L}_k, \overline{A}_k\right)$ can potentially take any value in $(-\infty, \infty)$, *(b)* $\Phi(x) = e^x$ and $E\left(Y_{k+1} \mid \overline{L}_k, \overline{A}_k\right)$ can potentially take any value in $(0, \infty)$, or *(c)* $\Phi(x) = e^x/(1+e^x)$ and Y_k is dichotomous.

REMARK 8.11. Furthermore, by Theorem 8.5 and Sec. 8.4, if *(a)*, *(b)*, or *(c)* is true, then the model in which γ^* is known (i.e., ψ in (8.44d) has but a single element) with all the other parameters left unrestricted is also a NPI model. Fully parametric likelihood inference based on the likelihood (8.47) for the unknown components of ρ in either of the above two models is available if we further restrict the unknown parameters to lie in finite dimensional parameter spaces.

REMARK 8.12. The likelihood function (8.47) can be written as $\mathcal{L}_1(\rho_1)\,\mathcal{L}_2(\rho_2)$ with $\psi \in \rho_1, \eta \in \rho_2$, ρ_1 and ρ_2 variation independent, and

$\rho = (\rho_1, \rho_2)$ if and only if the functions q_m are identically zero (i.e., there is no confounding). It follows that, in the presence of confounding, the treatment process $f\left[a_m \mid \bar{\ell}_m, \bar{a}_{m-1}; \eta\right]$ is no longer ancillary for ψ in the model with q_m known.

In Sec. 8.2.2 we considered semiparametric inference for ψ in the model with q_m (i.e., μ_1) known and ψ and η unknown finite dimensional parameters with $\Phi(x)$ either x or e^x. The following remark considers this model further.

REMARK 8.13. No reasonable semiparametric inference (e.g., based on g-estimation) is available for ψ in the semiparametric models with q_m known and ψ and η finite dimensional when Y_k is dichotomous and $\Phi(x) = e^x/(1+e^x)$ (because all influence functions for ψ depend on a high-dimensional smooth, i.e., a conditional expectation or density which is left unrestricted by the model). Specifically, it can be shown that although the semiparametric information bound for ψ is finite when $\Phi(x) = e^x/(1+e^x)$, the curse of dimensionality appropriate (CODA) semiparametric variance bound in the sense of Robins and Ritov (1997b) is zero. In contrast, with $\Phi(x) = x$ or e^x, the CODA bound is finite [as evidenced by the existence of our semiparametric g-estimators for ψ that we were able to calculate without smoothing].

REMARK 8.14. In the special case in which $K = 0$, things simplify considerably. To be concrete, suppose Y is dichotomous and $\Phi(x) = e^x/(1+e^x)$. From our previous definitions, we have the identity

$$\begin{aligned}\Phi^{-1}\{E[Y_1 \mid A_0, L_0]\} = \; & \gamma^*_{0,1}(\ell_0, a_0) + q_0(1, \ell_0, a_0) \\ & + \Phi^{-1}\left\{E\left[Y_{(0),1} \mid \ell_0, a_0 = 0\right]\right\} .\end{aligned} \tag{8.49}$$

This is natural to consider a model in which we assume q_0 is known (which is then varied in a sensitivity analysis), $\gamma^*_{0,1}(\ell_0, a_0)$ is known up to a finite dimensional parameter ψ, and the function $\Phi^{-1}\left\{E\left[Y_{(0),1} \mid \ell_0, a_0 = 0\right]\right\}$ of ℓ_0 is assumed known up to a finite dimensional parameter ω. Then we can jointly estimate ψ and ω by maximum likelihood by maximizing the conditional likelihood given A_0 and L_0. Note here, because of the non-nested structure (on account of K being zero), the conditional law of A_0 given L_0 and the marginal law of L_0 are ancillary for (ψ, ω).

For example, we could consider fitting, using an off-the-shelf logistic regression program, the model

$$logit\{pr[Y_1 = 1 \mid A_0, L_0]\} = \omega_0 + \omega_1' L_0 + \psi_0 A_0 + \psi_1' A_0 L_0 + \alpha A_0 + \alpha_1' L_0 A_0$$

where α_0 and α_1 are known fixed offsets (which are then varied in a sensitivity analysis) and $\omega_0, \omega_1, \psi_0, \psi_1$ are unknown finite dimensional parameters to be estimated. In this set-up and $q_0(1, L_0, A_0) \equiv \alpha_0 A_0 + \alpha_1' L_0 A_0$, $\gamma'_{0,1}(L_0, A_0) = \psi_0 A_0 + \psi_1' A_0 L_0$ and $\Phi^{-1}\{E[Y_{(0),1} \mid L_0, A_0 = 0]\} = \omega_0 + \omega_1' L_0$.

8.6. Marginal structural mean models. A marginal structural model is a model for the law of $\overline{Y}_{\overline{a},K+1}$ given a subset V_0^* of L_0 for all $\overline{a} \in \overline{\mathcal{A}}$. A marginal structural mean model specifies that

$$\Phi^{-1}\left\{E\left[Y_{\overline{a},k+1} \mid V_0^*\right]\right\} \equiv \gamma_{k+1}^*\left(\overline{a}_k, V_0^*\right) \tag{8.50}$$

where Φ is a known $1-1$ function and

$$\gamma_{k+1}^*\left(\overline{a}_k, V_0^*\right) \in \left\{\gamma_{k+1}^*\left(\overline{a}_k, V_0^*; \psi\right); \psi \in \boldsymbol{\psi}\right\} \tag{8.51}$$

and ψ is a (possibly infinite dimensional) unknown parameter taking values in $\boldsymbol{\psi}$. Our ultimate goal is to construct a sensitivity analysis for ψ.

8.6.1. Likelihood inference for marginal structural mean models. We shall construct the likelihood for our model based on the following decomposition.

$$\begin{aligned}
\Phi^{-1}\left\{E\left[Y_{\overline{a},k+1} \mid \overline{a}_k, \overline{\ell}_k\right]\right\} &= \sum_{m=0}^{k} \Phi^{-1}\left\{E\left[Y_{\overline{a},k+1} \mid \overline{a}_m, \overline{\ell}_m\right]\right\} \\
&- \Phi^{-1}\left\{E\left[Y_{\overline{a},k+1} \mid \overline{a}_{m-1}, \overline{\ell}_m\right]\right\} + \sum_{m=0}^{k} \Phi^{-1}\left\{E\left[Y_{\overline{a},k+1} \mid \overline{a}_{m-1}, \overline{\ell}_m\right]\right\} \qquad (8.52) \\
&- \Phi^{-1}\left\{E\left[Y_{\overline{a},k+1} \mid \overline{a}_{m-1}, \overline{\ell}_{m-1}, v_0^*\right]\right\} + \Phi^{-1}\left\{E\left[Y_{\overline{a},k+1} \mid v_0^*\right]\right\} \\
&\equiv \sum_{m=0}^{k} m_{k+1}\left(\overline{\ell}_m, \overline{a}_k\right) + m_{k+1}^*\left(\overline{\ell}_m, \overline{a}_k\right) + \gamma_{k+1}^*\left(\overline{a}_k, v_0^*\right) \qquad (8.53)
\end{aligned}$$

where

$$\begin{aligned}
m_{k+1}\left(\overline{\ell}_m, \overline{a}_k\right) = \; & \Phi^{-1}\left\{E\left[Y_{\overline{a},k+1} \mid \overline{a}_m, \overline{\ell}_m\right]\right\} \\
& - \Phi^{-1}\left\{E\left[Y_{\overline{a},k+1} \mid \overline{a}_{m-1}, \overline{\ell}_m\right]\right\}
\end{aligned} \tag{8.54}$$

and

$$\begin{aligned}
m_{k+1}^*\left(\overline{\ell}_m, \overline{a}_k\right) = \; & \Phi^{-1}\left\{E\left[Y_{\overline{a},k+1} \mid \overline{a}_{m-1}, \overline{\ell}_m\right]\right\} \\
& - \Phi^{-1}\left\{E\left[Y_{\overline{a},k+1} \mid \overline{a}_{m-1}, \overline{\ell}_{m-1}, v_0^*\right]\right\} .
\end{aligned} \tag{8.55}$$

However, to obtain an unrestricted variation-independent parameterization, we cannot directly parameterize $m_{k+1}\left(\overline{\ell}_m, \overline{a}_k\right)$ and $m_{k+1}^*\left(\overline{\ell}_m, \overline{a}_k\right)$. However, we can take the following approach. Define

$$\begin{aligned}
q_m\left(k+1, \overline{\ell}_m, \overline{a}_k, a_m^*\right) \equiv \; & \Phi^{-1}\left\{E\left(Y_{\overline{a},k+1} \mid \overline{\ell}_m, \overline{a}_m\right)\right\} \\
& - \Phi^{-1}\left\{E\left(Y_{\overline{a},k+1} \mid \overline{\ell}_m, \overline{a}_{m-1}, a_m^*\right)\right\}
\end{aligned} \tag{8.56}$$

and

$$\nu_m^*\left(k+1,\bar{\ell}_m,\bar{a}_k\right)=\Phi^{-1}\left\{E\left(Y_{\bar{a},k+1}\mid\bar{\ell}_m,\bar{a}_{m-1}\right)\right\} - \Phi^{-1}\left\{E\left(Y_{\bar{a},k+1}\mid\bar{\ell}_{m-1},\bar{a}_{m-1},\ell_m=0\right)\right\},\ m>0 \tag{8.57a}$$

where $\ell_m = 0$ is a baseline level of ℓ_m and

$$\nu_m^*\left(k+1,\bar{\ell}_m,\bar{a}_k\right)=\Phi^{-1}\left\{E\left(Y_{\bar{a},k+1}\mid\bar{\ell}_m,\bar{a}_{m-1}\right)\right\} - \Phi^{-1}\left\{E\left(Y_{\bar{a},k+1}\mid v_0^*,\ell_0\backslash v_0^*=0\right)\right\},\ m=0 \tag{8.57b}$$

where $L_0\backslash V_0^*$ are those components of L_0 other than V_0^*.

We thus obtain that for $\Phi(x) = x$ or e^x

$$m_{k+1}\left(\bar{\ell}_m,\bar{a}_k\right)= - \Phi^{-1}\int\Phi\left[-q_m\left(k+1,\bar{\ell}_m,\bar{a}_k,a_m^*\right)\right]dF\left[a_m^*\mid\bar{\ell}_m,\bar{a}_{m-1}\right]. \tag{8.58}$$

When $\Phi(x) = e^x/(1+e^x)$,

$$m_{k+1}\left(\bar{\ell}_m,\bar{a}_k\right)=\Phi^{-1}(p)-\Phi^{-1}\left\{E\left(Y_{\bar{a},k+1}\mid\bar{\ell}_m,\bar{a}_{m-1}\right)\right\} \tag{8.59}$$

where $p = p\left(k+1,\bar{\ell}_m,\bar{a}_k\right)$ is the unique solution to

$$E\left(Y_{\bar{a},k+1}\mid\bar{\ell}_m,\bar{a}_{m-1}\right)= \int\left\{1+(1-p)p^{-1}\exp\left[q_m(k+1,\bar{\ell}_m,\bar{a}_k,a_m^*)\right]\right\}^{-1}dF\left(a_m^*\mid\bar{\ell}_m,\bar{a}_{m-1}\right). \tag{8.60}$$

Furthermore, we have

$$m_{k+1}^*\left(\bar{\ell}_m,\bar{a}_k\right)=\nu_m^*\left(k+1,\bar{\ell}_m,\bar{a}_k\right)-\Gamma^*\left(k+1,\bar{\ell}_{m-1},\bar{a}_k,v_0^*\right) \tag{8.61}$$

where

$$\Gamma^*\left(k+1,\bar{\ell}_{m-1},\bar{a}_k,v_0^*\right)=-\Phi^{-1}\left[E\left(Y_{\bar{a},k+1}\mid\bar{\ell}_{m-1},\ell_m=0,\bar{a}_{m-1}\right)\right] + \Phi^{-1}\left[E\left(Y_{\bar{a},k+1}\mid\bar{\ell}_{m-1},\bar{a}_{m-1}\right)\right],\ m>0 \tag{8.62a}$$

and

$$\Gamma^*\left(k+1,\bar{\ell}_{m-1},\bar{a}_k,v_0^*\right)=-\Phi^{-1}\left[E\left(Y_{\bar{a},k+1}\mid v_0^*,\ell_0\backslash v_0^*=0\right)\right] + \Phi^{-1}\left[E\left(Y_{\bar{a},k+1}\mid v_0^*\right)\right],\ m=0\,. \tag{8.62b}$$

For $\Phi(x) = x$ or e^x

$$\Gamma^*\left(k+1,\bar{\ell}_{m-1},\bar{a}_k,v_0^*\right)= \Phi^{-1}\left[\int\Phi\left[\nu_m^*\left(k+1,\bar{\ell}_m,\bar{a}_k\right)\right]dF\left[\ell_m\mid\bar{\ell}_{m-1},\bar{a}_{m-1},v_0^*\right]\right], \tag{8.63}$$

while, for $\Phi(x) = e^x/(1+e^x)$, $\exp\{-\Gamma^*(k+1, \bar{\ell}_{m-1}, \bar{a}_k, v_0^*)\}$ is the unique solution x to

$$(1-p)^2 = \int [(1-p)\nu_m^*(k+1, \bar{\ell}_m, \bar{a}_k)x^{-1} + p]^{-1} dF[\ell_m \mid \bar{\ell}_{m-1}, \bar{a}_{m-1}, v_0^*] \tag{8.64}$$

with

$$p \equiv E[Y_{\bar{a},k+1} \mid \bar{\ell}_{m-1}, \bar{a}_{m-1}, v_0^*] \ . \tag{8.65}$$

To construct the likelihood function for non-parametric, semiparametric, and/or parametric versions of our marginal structural mean model, we consider again the sets of densities (8.44a)–(8.44c). We will also consider the following sets of functions.

$$\{\gamma_{k+1}^*(\bar{a}_k, v_0^*; \psi); \psi \in \boldsymbol{\psi}\} \tag{8.66a}$$

$$\{q_m(k+1, \bar{\ell}_m, \bar{a}_k, a_m^*; \mu_1);\ q_m(k+1, \bar{\ell}_m, \bar{a}_k, a_m) = 0 \ and \ \mu_1 \in \boldsymbol{\mu}_1\} \tag{8.66b}$$

and

$$\left\{\nu_m^*(k+1, \bar{\ell}_m, \bar{a}_k; \mu_2);\ \nu_m^*(k+1, \bar{\ell}_{m-1}, v_0^*, \ell_m \backslash v_0^* = 0, \bar{a}_k; \mu_2) = 0 \ and \ \mu_2 \in \boldsymbol{\mu}_2\right\} . \tag{8.66c}$$

$$\{s(v_0^*; \omega); \omega \in \boldsymbol{\omega}\} \ . \tag{8.66d}$$

In (8.66c), we set $\ell_m \backslash v_0^*$ to be ℓ_m if $m \neq 0$. Then the likelihood function can be written again as (8.47) with $\varepsilon_{k+1}(\rho)$ as in (8.48), $\rho = (\psi, \eta, \gamma, \mu, \omega)$ with variation-independent components and ψ, μ, and ω as redefined in (8.66), with $E[Y_0 \mid \overline{L}_{-1}, \overline{A}_{-1}, V_0^*; \rho] = s(V_0^*; \omega)$, and $E[Y_{k+1} \mid \overline{L}_k, \overline{A}_k; \rho]$ for $k \geq 0$ given by the RHS of (8.53), except with the parameterized versions of $m_{k+1}(\bar{\ell}_m, \bar{a}_k)$, $m_{k+1}^*(\bar{\ell}_m, \bar{a}_k)$, and $\gamma_{k+1}^*(\bar{a}_k, v_0^*)$ replacing the unparameterized versions. Note by (8.58), (8.59), (8.61), (8.62), (8.63), and (8.64) $m_{k+1}(\bar{\ell}_m, \bar{a}_k)$ and $m_{k+1}^*(\bar{\ell}_m, \bar{a}_k)$ can be parameterized in terms of the parameter ρ. Again, there is no parameter γ_1 when Y_k is dichotomous and $\Phi(x) = e^x/(1+e^x)$. The components of ρ will be variation independent under the same conditions described in Remark 8.9.

REMARK 8.15. With the MSM-specific redefinitions of the various quantities, Remarks 8.9–8.13 of Sec. 8.5 continue to hold for MSMs. In the next section, we provide a semiparametric estimator for the parameter ψ of a marginal structural mean model with $\Phi(x) = x$ or e^x when the parameter η is finite-dimensional.

REMARK 8.16. The sharp null hypothesis that

$$\overline{Y}_{\overline{a},K+1} = \overline{Y}_{\overline{a}^*,K+1} \text{ w.p.1 for all } \overline{a} \text{ and } \overline{a}^* \tag{8.67}$$

implies

$$\gamma^*_{k+1}\left(\overline{a}_k, v^*_0\right) \text{ does not depend on } \overline{a}_k \tag{8.68a}$$

$$\nu^*_m\left(k+1, \overline{\ell}_m, \overline{a}_k\right) \text{ depends on } \overline{a}_k \text{ only through } \overline{a}_{m-1}\ . \tag{8.68b}$$

$$\begin{gathered} q_m\left(k+1, \overline{\ell}_m, \overline{a}_k, a^*_m\right) \equiv q_m\left(k+1, \overline{\ell}_m, \{\overline{a}_{m-1}, a_m\}, a^*_m\right) \\ \text{depends on } \overline{a}_k \text{ only through } \overline{a}_m \end{gathered} \tag{8.68c}$$

$$\begin{gathered} q_m\left(k+1, \overline{\ell}_m, \{\overline{a}_{m-1}, a_m\}, a^*_m\right) = -q_m\left(k+1, \overline{\ell}_m, \{\overline{a}_{m-1}, a^*_m\}, a_m\right) \\ \text{is anti-symmetric in } a_m \text{ and } a^*_m\ . \end{gathered} \tag{8.68d}$$

The above implies that, in contrast with SNMs, if we are performing a sensitivity analysis in which we fix the non-identifiable function q_m, then, in order to test the null hypothesis (8.67) it is necessary for us to restrict our choice of q_m. In particular, our choice of q_m must satisfy (8.68c) and (8.68d), which can be shown to be the only restrictions on q_m implied by (8.67).

Restrictions (8.68c) and (8.68d) together are equivalent to the condition that $r_m\left(k+1, \overline{\ell}_m, \overline{a}_k, a^*_m\right) = 0$ where $r_m\left(k+1, \overline{\ell}_m, \overline{a}_k, a^*_m\right) \equiv q_m\left(k+1, \overline{\ell}_m, \overline{a}_k, a^*_m\right) - q_m\left(k+1, \overline{\ell}_m\left(\overline{a}_{m-1}, 0\right), a^*_m\right) + q_m\left(k+1, \overline{\ell}_m, \left(\overline{a}_{m-1}, 0\right), a_m\right)$. Note, with $\Phi(x) = x$,

$$\begin{aligned} & r_m\left(k+1, \overline{\ell}_m, \overline{a}_k, a^*_m\right) \\ & \quad = \left\{E\left[Y_{\overline{a},k+1} \mid \overline{\ell}_m, \overline{a}_m\right] - E\left[Y_{(\overline{a}_{m-1},0),k+1} \mid \overline{\ell}_m, \overline{a}_m\right]\right\} \\ & \qquad - \left\{E\left[Y_{\overline{a},k+1} \mid \overline{\ell}_m, \overline{a}_{m-1}, a^*_m\right] - E\left[Y_{(\overline{a}_{m-1},0),k+1} \mid \overline{\ell}_m, \overline{a}_{m-1}, a^*_m\right]\right\} \end{aligned}$$

which has an interpretation as a type of current-treatment interaction function.

It would be unlikely to hold prior beliefs that (8.68a) is true but (8.67) is false. This implies it would be unlikely to hold prior beliefs that (8.68a) is true, but (8.68c) and/or (8.68d) were false. Thus, in conducting a sensitivity analysis which treats the selection bias function $q_m\left(k+1, \overline{\ell}_m, \overline{a}_k, a^*_m\right)$ as known and then tests (8.68a) from the data, it would be important to choose $q_m\left(k+1, \overline{\ell}_m, \overline{a}_k, a^*_m\right)$ to satisfy (8.68c) and (8.68d), which can be accomplish by setting $r_m\left(k+1, \overline{\ell}_m, \overline{a}_k, a^*_m\right) = 0$ and then choosing the single function $q_m\left(k+1, \overline{\ell}_m, \left(\overline{a}_{m-1}, 0\right), a_m\right)$. It follows that this same

approach to choosing $q_m\left(k+1, \bar{\ell}_m, \bar{a}_k, a_m^*\right)$ should be used whenever one wishes to test the hypothesis (8.67).

REMARK 8.17. **Direct effect null hypothesis:** Suppose that $\overline{A}_K = \left(\overline{A}_{PK}, \overline{A}_{ZK}\right)$ where P and Z refer to two different treatments. The sharp null hypothesis that $\overline{A}_{PK}$ has no direct effect on $\overline{Y}_{K+1}$ when treatment $\overline{A}_{ZK}$ is set to a particular value $\bar{a}_{ZK}$ is the hypothesis

$$\overline{Y}_{(\bar{a}_P, \bar{a}_Z)} = \overline{Y}_{\left(\bar{a}_P^*, \bar{a}_Z\right)} \text{ w.p.1 for all } \bar{a}_P, \bar{a}_P^*, \bar{a}_Z^* \ . \tag{8.69}$$

This implies that

$$\gamma_{k+1}^*\left(\bar{a}_k, v_0^*\right) = \gamma_{k+1}^*\left(\bar{a}_{Zk}, v_0^*\right) \tag{8.70}$$

$$\nu_m^*\left(k+1, \bar{\ell}_m, \bar{a}_k\right) = \nu_m^*\left(k+1, \bar{\ell}_m, \left(\bar{a}_{m-1}, \bar{a}_{Zk}\right)\right) \tag{8.71}$$

$$q_m^*\left(k+1, \bar{\ell}_m, \bar{a}_k, a_m^*\right) = q_m^*\left(k+1, \bar{\ell}_m, \left\{\bar{a}_{Zk}, \bar{a}_{Pm}\right\}, a_m^*\right) \tag{8.72}$$

and

$$\begin{aligned} &q_m^*\left(k+1, \bar{\ell}_m, \left(\bar{a}_{Zk}, \bar{a}_{Pm}\right), \left(a_{Zm}, a_{Pm}^*\right)\right) \\ &\quad = -q_m^*\left(k+1, \bar{\ell}_m, \left(\bar{a}_{Zk}, \bar{a}_{P(m-1)}, a_{Pm}^*\right), \left(a_{Zm}, a_{Pm}\right)\right) \ . \end{aligned} \tag{8.73}$$

It is important to note in (8.73) that there is no a_{Zm}^* in addition to a_{Zm}.

REMARK 8.18. **Inadequacies of MSMs and SNMs for sensitivity analysis in randomized trials with non-compliance:** Consider a randomized trial with all or none compliance in which the observed data are $(A_0, A_1, Y = Y_2)$, all of which are dichotomous. $A_0 \equiv A_{P0}$ is a randomization indicator, $A_1 \equiv A_{Z1}$ is the actual treatment, and $Y = Y_2$ is the observed outcome. We make the exclusion restriction that A_0 has no direct effect on Y when A_1 is fixed, i.e.,

$$Y_{(0,a_1)} = Y_{(1,a_1)} \equiv Y_{a_1} \text{ w.p.1 for } a_1 \in \{0,1\} \ . \tag{8.74}$$

Further, since A_0 is randomized, we assume

$$A_0 \coprod \left(Y_1, Y_0\right) \ . \tag{8.75}$$

Now, given data $(A_0, A_1, Y \equiv Y_{A_1})$ all dichotomous, a marginal structural mean model with $\Phi(x) = e^x/(1+e^x)$ has 15 parameters: $f(a_0, a_1)$ has 3 parameters, and $\gamma_1^*(a_0, a_1)$, $q_0\left[\left(a_0, a_1\right), a_0^* = (1-a_0)\right]$, and $q_1\left[\left(a_0, a_1\right), a_1^* = (1-a_1)\right]$ have 4 each. Now our restrictions (8.74) and (8.75) imply that

$$\gamma_1^*(a_0, a_1) = \gamma_1^*(1-a_0, a_1) \equiv \gamma_1^*(a_1) \tag{8.76}$$

and

$$q_0\left[(a_0, a_1), a_0^* = (1 - a_0)\right] = 0 . \tag{8.77}$$

However, in the absence of knowledge of the distribution of the observed data F_O, these restrictions (8.74)–(8.75) place no other functional equality constraints on the remaining 9 parameters $\gamma_1^*(a_1)$, $q_1\big((a_0, a_1), a_1^* = 1 - a_1\big)$, and $f(a_0, a_1)$. However, (8.74) plus the "marginal" randomization assumption

$$A_0 \coprod Y_{a_1}; a_1 \in \{0, 1\} \tag{8.78}$$

also imply (8.76) and (8.77). Now suppose we are given the distribution F_O of $O = (A_0, A_1, Y)$. This distribution has 7 parameters, so if we impose (8.76) and (8.77), we would not generally expect the parameter of interest $\gamma_1^*(a_1)$ to be identified. Now since (8.74) and (8.78) imply (8.76) and (8.77), the so-called natural bounds of Robins (1989) and Manski (1990) on $E(Y_1) - E(Y_0) = \gamma_1^*(1) - \gamma_1^*(0)$ determined by F_O, (8.74) and (8.78) will have width less than or equal to the bounds determined by F_O, (8.76), and (8.77). However, Balke and Pearl (1997) shows the natural bounds can, for certain F_O, be strictly wider than the bounds determined by F_O, (8.74) and (8.75). Thus, it follows that for such an F_O, (8.74) and (8.75) imply additional constraints on the parameters of our MSM beyond (8.76) and (8.77). However, it is not easy to write down these additional constraints implying that, in general, the use of our MSM parameterization is not particularly convenient for analyzing a randomized trial with all or none compliance. [Indeed, Balke and Pearl (1997) show that, given (8.74) and (8.75), for certain special F_O, the joint distribution of (Y_0, Y_1, A_0, A_1) is identified and thus $\gamma_1^*(a_1)$ and $q_1[(a_0, a_1), a_1^* = 1 - a_1]$ are precisely known.] It can be shown similarly that our logistic SNM parameterization is also inconvenient in the same sense. However, both our MSM and our logistic SNM parameterizations should be suitable for analyzing most observational studies or randomized equivalence trials with non-compliance since *(i)* as argued in Robins (1997b), in observational studies, practicing epidemiologists give zero prior probability to the event that (8.78) or the event (8.74) are true, and *(ii)* Robins (1998c) shows that randomized equivalence trials with non-compliance, that compare a known active therapy to a new therapy, can be viewed statistically as to observational studies. An exception would be observational studies in which one thinks there may be a "near instrument" such as studies of the effect of education on schooling where the month in which a student dropped out of high school is considered an instrument.

REMARK 8.19. In our parameterization of MSMs, we could have replaced the parameter $q_m\left(k+1, \bar{\ell}_m, \bar{a}_k, a_m^*\right)$ of (8.56) by

$$\begin{aligned} &q_m\left(k+1, \bar{\ell}_m, \bar{a}_k\right) \equiv \\ &\Phi^{-1}\left\{E(Y_{\bar{a},k+1} \mid \bar{\ell}_m, \bar{a}_m)\right\} - \Phi^{-1}\left\{E(Y_{\bar{a},k+1} \mid \bar{\ell}_m, \bar{a}_{m-1}, A_m \neq a_m)\right\}. \end{aligned} \tag{8.79}$$

With this parameterization, the marginal structural mean model with $\Phi(x) = e^x/(1+e^x)$ is a special case of the selection odds model of Sec. 7 and thus remains, based on the results of that section, a NPI model. In addition, as noted in that section, if A_m is a continuous random variable and $pr\left[A_m = a_m \mid \bar{\ell}_m, \bar{a}_{m-1}\right] = 0$ for all a_m, then the density $f\left(a_m \mid \bar{\ell}_m, \bar{a}_{m-1}\right)$ remains ancillary for the parameter ψ. That is, Remark 8.12 of Sec. 8.5 holds for all functions $q_m\left(k+1, \bar{\ell}_m, \bar{a}_k\right)$, rather than just holding when the function is identically zero (i.e., when there is no confounding). It also follows that the MSM parameterization described in this section can be used as a parameterization of our selection odds model of Sec. 7, for purposes of fully parametric likelihood-based inference.

Unfortunately, as we now describe, there is a major difficulty with using the parameterization discussed in this remark if we wish to test the null hypothesis (8.67). Therefore, the approach described in the last paragraph of Remark 8.16 of this subsection should be used. The null hypothesis (8.67) implies not only that $q_m\left(k+1, \bar{\ell}_m, \bar{a}_k\right)$ is a function of $\bar{a}_k$ only through $\bar{a}_m$, but it also implies additional joint restrictions on $q_m\left(k+1, \bar{\ell}_m, \bar{a}_k\right)$ and the identifiable density $f\left(a_m \mid \bar{\ell}_m, \bar{a}_{m-1}\right)$. Suppose, therefore, that one takes the point of view that it is *a priori* unlikely that (8.68a) is true but (8.67) is false (although this is logically possible). Then it is not appropriate to simply use the parameterization $q_m\left(k+1, \bar{\ell}_m, \bar{a}_k\right)$ of (8.79) for the selection bias function in an estimation procedure which treats this function as known and then tests (8.68a) from the data, as there is no guarantee that the joint restriction on the chosen $q_m\left(k+1, \bar{\ell}_m, \bar{a}_k\right)$ and $f\left(a_m \mid \bar{\ell}_m, \bar{a}_{m-1}\right)$ implied by (8.67) will be fulfilled.

8.6.2. Semiparametric inference in marginal structural mean models. We consider a marginal structural mean model in which

(8.80a) $q_m\left(k+1, \overline{L}_m, \overline{A}_k, a_m^*\right)$ *is known*

(8.80b) ψ *in (8.66a) is a finite dimensional vector with true value* ψ^*

(8.80c) η *in (8.44a) is a finite dimensional vector with true value* η^* .

Let $g_k\left(\overline{A}_k, V_0^*\right)$ and $s_k\left(\overline{A}_k, \overline{L}_k\right)$ be, respectively, user-specified $\dim\psi + \dim\eta$ and $\dim\eta$ vector-valued functions. Let

$$W^\dagger(\eta) = W^\dagger(\eta, s) = \tag{8.81}$$
$$\left(0', \left[\sum_{k=0}^{K} s_k\left(\overline{A}_k, \overline{L}_k\right) - \int s\left(\overline{A}_k, \overline{L}_k\right) dF\left(A_k \mid \overline{A}_{k-1}, \overline{L}_k; \eta\right)\right]'\right)'$$

where 0 is a $\dim\psi$ vector of 0's. Let $g\left(\overline{A}, V_0^*\right)$ be the $(\dim\psi + \dim\eta) \times (K+1)$ matrix-valued function with columns $g_k\left(\overline{A}_k, V_0^*\right), k = 0, \ldots, K$.

Let

$$H_k(\psi,\eta) = \left\{ Y_{k+1} - \sum_{m=0}^{k} \int q_m\left(k+1, \overline{L}_m, \overline{A}_k, a_m^*\right)\, dF\left(a_m^* \mid \overline{L}_m, \overline{A}_{m-1}; \eta\right) - \gamma_{k+1}^*\left(\overline{A}_k, V_0^*; \psi\right) \right\} \Big/ \left\{ \prod_{m=0}^{k} f\left[A_m \mid \overline{A}_{m-1}, \overline{L}_m\right] \right\}, \tag{8.82a}$$

for $\Phi(x) = x$

$$H_k(\psi,\eta) = \left\{ Y_{k+1} \prod_{m=0}^{k} \int \exp\left\{-q_m(k+1, \overline{L}_m, \overline{A}_k, a_m^*)\right\} dF(a_m^* \mid \overline{L}_m, \overline{A}_{m-1}; \eta) - \exp\left\{\gamma_{k+1}^*\left(\overline{A}_k, V_0^*; \psi\right)\right\} \right\} \Big/ \left\{ \prod_{m=0}^{k} f\left[A_m \mid \overline{A}_{m-1}, \overline{L}_m\right] \right\}, \tag{8.82b}$$

for $\Phi(x) = e^x$.

Let

$$W(\psi,\eta) \equiv W(\psi,\eta,g) = g\left(\overline{A}, V_0^*\right)\left[H_0(\psi,\eta), \ldots, H_K(\psi,\eta)\right]' .$$

THEOREM 8.14. *In the semiparametric marginal structural mean model characterized by (8.80) if* $\Phi(x) = x$ *or* e^x*, then, subject to regularity conditions,*
(a). $E\left[W\left(\psi^*, \eta^*\right)\right] = 0$*, and* $E\left[W^{\dagger}\left(\eta^*\right)\right] = 0$*;*
(b). $\left(\widehat{\psi}, \widehat{\eta}\right) = \left(\widehat{\psi}(g,s), \widehat{\eta}(g,s)\right)$ *solving*

$$0 = \sum_i W_i(\psi,\eta) + W_i^{\dagger}(\eta) \tag{8.83}$$

is a consistent asymptotically normal estimator of (ψ^*, η^*)*;*
(c). $\left\{W(\psi,\eta;g) + W^{\dagger}(\eta,s)\right\}$*, as g and s vary, is the orthogonal complement to the nuisance tangent space for the model; (d). There exists* g_{eff}, s_{eff} *such that the asymptotic variance of* $\left[\widehat{\psi}(g_{eff}, s_{eff}), \widehat{\eta}(g_{eff}, s_{eff})\right]$ *attains the semiparametric variance bound for the model. The crucial result 1 is proved in Appendix B.*

REMARK 8.20. An alternative estimation procedure is to first estimate η by the partial likelihood estimator $\widehat{\eta}$ maximizing $\prod_{i=1}^{n} PL_i(\eta)$ where

$$PL(\eta) = \prod_{m=0}^{K} f\left[A_m \mid \overline{L}_m, \overline{A}_{m-1}; \eta\right] . \tag{8.84}$$

We then estimate ψ by the solution to $0 = \sum_i W_i(\psi, \widehat{\eta}) = 0$ with each $g_k(\cdot,\cdot)$ now a $\dim\psi$ vector-valued function. This class will not contain a semiparametric efficient estimator. However, this class avoids the need to compute the integral in (8.81).

REMARK 8.21. The integral in (8.82a) and (8.82b) can be difficult to compute. The need to compute the integral can be avoided in two different ways. The first way is to change the model by replacing the known $q\left(k+1, \overline{L}_m, \overline{A}_k, a_m^*\right)$ by $q_m\left(k+1, \overline{L}_m, \overline{A}_k\right)$ of Eq. (8.79). Then, for example, in defining $H_k(\psi,\eta)$, we would replace the integral in (8.81a) by $\left\{f\left(A_m \mid \overline{L}_m, \overline{A}_{m-1}\right) + \left[1 - f\left(A_m \mid \overline{L}_m, \overline{A}_{m-1}\right)\right] q_m\left(k+1, \overline{L}_m, \overline{A}_k\right)\right\}$ if $pr\left[A_m = a_m \mid \overline{L}_m, \overline{A}_{m-1}\right] \neq 0$ for $a_m = A_m$ and by $q_m\left(k+1, \overline{L}_m, \overline{A}_k\right)$ if $pr\left[A_m = a_m \mid \overline{L}_m, \overline{A}_{m-1}\right] = 0$ for $a_m = A_m$. In this latter case, the partial likelihood estimator of η is efficient, $f\left(A_m \mid \overline{L}_m, \overline{A}_{m-1}\right)$ is ancillary for ψ, and the class of estimators in Remark 8.20 will include an efficient estimator.

A second approach to avoiding the integrals in (8.82a) and (8.92b) is to sample. Specifically, suppose $\Phi(x) = x$. Then we redefine $H_k(\psi,\eta)$ in (8.82a) as

$$H_k(\psi,\eta) = Y_{k+1} - J^{-1}\sum_{j=1}^{J}\sum_{m=0}^{k} q_m\left(k+1, \overline{L}_m, \overline{A}_k, a_m^{*j}\right) - \gamma_{k+1}^*\left(\overline{A}_k, V_0^*; \psi\right)$$

where the a_m^{*j} are independent draws from $f\left[a_m \mid \overline{L}_m, \overline{A}_{m-1}; \eta\right]$. A related algorithm is available when $\Phi(x) = e^x$.

REMARK 8.22. Note that our estimator of ψ under the model (8.80) did not require that we compute a "non-parametric smooth" (i.e., an estimate of conditional expectation or density whose functional form is left unrestricted by the model (8.80)). In contrast, it can be shown that to estimate model ψ in model (8.80) with $\Phi(x) = e^x/(1+e^x)$, we would have to estimate such a "non-parametric smooth," which is not feasible in moderate sized samples due to the curse of dimensionality. Formally, this can be expressed by the fact that the curse of dimensionality-appropriate information bound is zero for ψ in model (8.80) when $\Phi(x) = e^x/(1+e^x)$ but is positive when $\Phi(x) = x$ or e^x. The reason for this is that the expectation operator E is a linear operator and thus commutes with addition (i.e., $\Phi(x) = x$) and multiplication ($\Phi(x) = e^x$).

REMARK 8.23. A continuous time version of a marginal structural model in which the treatment process can jump in continuous time is as follows. For simplicity, we consider an outcome $Y = Y(\tau)$ with counterfactuals $Y_{\overline{a}}, \overline{a} = \{a(u); 0 \leq u \leq \tau\} \in \overline{\mathcal{A}}$. Here, τ is a fixed, non-random end-of-follow-up time. We assume

$$\Phi^{-1}\left[E\left(Y_{\overline{a}} \mid V_0^*\right)\right] = \gamma^*\left(\overline{a}, V_0^*\right) \tag{8.85}$$

for a known continuous increasing function Φ. Further we assume

$$\gamma^*(\bar{a}, V_0^*) \in \{\gamma^*(\bar{a}, V_0^*; \psi); \psi \in \boldsymbol{\psi}\} . \tag{8.86}$$

We assume the processes $\overline{L}(u)$ and $\overline{A}(u)$ have CADLAG sample paths, and we let $L(t^-)$ and $A(t^-)$ be the left-hand limits of $A(u)$ and $L(u)$ as $u \uparrow t$. Further, for simplicity, we assume $A(t) \in \{0, 1\}$ and that the hazard (intensity)

$$\lambda_A\left(t \mid \overline{L}\left(t^-\right), \overline{A}\left(t^-\right)\right) = \lim_{h \to 0} h^{-1} pr\left[A(t+h) \neq A\left(t^-\right) \mid \overline{L}\left(t^-\right), \overline{A}\left(t^-\right)\right]$$

is uniformly bounded and smooth as a function of t with probability 1. Let

$$\begin{aligned} q\left(t, \bar{\ell}\left(t^-\right), \bar{a}\right) = {} & \Phi^{-1}\left\{E\left(Y_{\bar{a}} \mid \bar{\ell}\left(t^-\right), \bar{a}(t)\right)\right\} \\ & - \Phi^{-1}\left\{E\left(Y_{\bar{a}} \mid \bar{\ell}\left(t^-\right), \bar{a}\left(t^-\right), A(t) \neq a(t)\right)\right\} . \end{aligned} \tag{8.87}$$

If $\Phi(x) = x$, let

$$\begin{aligned} H(\psi) \equiv {} & \Bigg\{ Y - \sum_{\{u; A(u) \neq A(u^-)\}} q\left(u, \overline{L}(u^-), \overline{A}\right) \\ & - \int_0^\tau q\left(t, \overline{L}(t^-), \overline{A}\right) \lambda_A\left(t \mid \overline{L}(t^-), \overline{A}(t^-)\right) dt - \gamma^*(\overline{A}, V_0^*; \psi) \Bigg\} \Bigg/ \\ & \left\{ \prod_{\{u; A(u) \neq A(u^-)\}} \lambda_A\left[u \mid \overline{L}(u^-), \overline{A}(u^-)\right] \exp\left[-\int_0^\tau \lambda_A\left(t \mid \overline{L}(t^-), \overline{A}(t^-)\right) dt\right] \right\} . \end{aligned} \tag{8.88}$$

If $\Phi(x) = e^x$, let

$$\begin{aligned} H(\psi) = {} & \\ & \left\{ Y \exp\left[- \sum_{\{u; A(u) \neq A(u^-)\}} q\left(u, \overline{L}\left(u^-\right), \overline{A}\right) + z\left(\overline{L}, \overline{A}\right) \right] - \gamma^*\left(\overline{A}, V_0^*; \psi\right) \right\} \Bigg/ \\ & \left\{ \prod_{\{u; A(u) \neq A(u^-)\}} \lambda_A\left[u \mid \overline{L}(u^-), \overline{A}(u^-)\right] \exp\left[-\int_0^\tau \lambda_A\left(t \mid \overline{L}(t^-), \overline{A}(t^-)\right) dt\right] \right\} , \end{aligned} \tag{8.89}$$

where

$$z\left(\overline{L}, \overline{A}\right) \equiv \int_0^\tau \left\{\exp\left[-q\left(t, \overline{L}\left(t^-\right), \overline{A}\right)\right] - 1\right\} \lambda_A\left(t \mid \overline{L}\left(t^-\right), \overline{A}\left(t^-\right)\right) dt .$$

Let $W(\psi) \equiv W(\psi, g) = g\left(\overline{A}, V_0^*\right) H(\psi)$ where $g\left(\bar{a}, V_0^*\right)$ is a function chosen by the investigator, usually chosen to be of dimension of ψ when ψ is finite-dimensional. Our main result is the following.

THEOREM 8.15. *$E[W(\psi^*)] = 0$ where ψ^* is the true value of ψ in (8.86).*

Proof. It follows from Theorem (8.14a) by letting the time δt between measurements at k and $k+1$ go to zero. It also can be shown

directly by the continuous time version of the proof of Theorem (8.14a) in Appendix B. □

In practice, in order to estimate ψ^*, we will in general specify a model such as the Cox proportional hazards model

$$\lambda_A\left(t \mid \overline{L}\left(t^-\right), \overline{A}\left(t^-\right)\right) = \lambda_0(t) \exp\left(\eta_0' H(t)\right)$$

where $H(t)$ is a known vector-valued function of $\left(\overline{L}\left(t^-\right), \overline{A}\left(t^-\right)\right)$ and $\lambda_0(t)$ is an unspecified positive function, and η_0 is an unknown parameter vector to be estimated. We then estimate ψ by solving the equation $\sum_i W_i(\psi) = 0$ with $\lambda_A\left(t \mid \overline{A}\left(t^-\right), \overline{L}\left(t^-\right)\right)$ replaced by its estimate based on the above Cox model.

8.7. Marginal structural distribution models. In this section, for ease of presentation, we only consider univariate marginal structural distribution models. The generalization to include multivariate marginal structural distribution models is straightforward. A univariate marginal structural distribution model specifies a model for a counterfactual continuous outcome $Y_{\overline{a}} \equiv Y_{\overline{a}(K+1)}$ measured at end of follow-up as a function of baseline variables V_0^*. In this setting, we have the observed data $O = \left(\overline{L}, \overline{A}, Y\right) = \left(\overline{L}_K, \overline{A}_K, Y_{K+1}\right)$. Then our model states

$$f_{Y_{\overline{a}}}\left(y \mid V_0^*\right) \in \left\{f\left(y \mid V_0^*; \psi\right); \psi \in \boldsymbol{\psi}\right\} \tag{8.90}$$

where $f\left(y \mid V_0^*; \psi\right)$ is a known density depending on an unknown parameter (possibly infinite dimensional) ψ.

8.7.1. Likelihood inference for marginal structural distribution models. We shall consider the likelihood based on the following parameterization. This parameterization is useful for the semiparametric sensitivity analysis estimators described in Sec. 8.7.2 below. However, for fully parametric likelihood-based inference, as we shall see, this parameterization results in essentially an intractable likelihood function.

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

$$pr\left[m\left(Y_{\overline{a}}, \overline{\ell}_m, \overline{a}\right) < y \mid \overline{\ell}_m, \overline{a}_m\right] = pr\left(Y_{\overline{a}} < y \mid \overline{\ell}_m, \overline{a}_{m-1}\right) . \tag{8.91}$$

Let $m^*\left(y, \overline{\ell}_m, \overline{a}\right)$ be the unique increasing function of y satisfying

$$pr\left[m^*\left(Y_{\overline{a}}, \overline{\ell}_m, \overline{a}\right) < y \mid \overline{\ell}_m, \overline{a}_{m-1}\right] = pr\left[Y_{\overline{a}} < y \mid \overline{\ell}_{m-1}, \overline{a}_{m-1}, v_0^*\right] . \tag{8.92}$$

Set

$$M_{K+1}^* = Y$$

for $k = K, K-1, \ldots, 0$, set

$$M_k = m\left(M_{k+1}^*, \overline{L}_m, \overline{A}\right) \tag{8.93}$$

and

$$M_k^* = m^*\left(M_k, \overline{L}_m, \overline{A}\right) . \tag{8.94}$$

By the proof of Theorem 11.1 in Robins (1999), we have

$$M_0^* \coprod \overline{L}_K \mid \overline{A}_K \tag{8.95}$$

$$f_{M_0^*}\left(y \mid \overline{A}_k, V_0^*\right) = f_{Y_{\overline{a}}}\left(y \mid V_0^*\right) \tag{8.96}$$

$$f(O) = \left\{\partial M_0^*/\partial Y\right\} f\left[M_0^* \mid \overline{A}_k, V_0^*\right] f\left(\overline{L}_K, \overline{A}_K\right) . \tag{8.97}$$

However, to obtain an unrestricted variation-independent parameterization, we cannot directly parameterize $m\left(y, \overline{\ell}_m, \overline{a}\right)$ and $m^*\left(y, \overline{\ell}_m, \overline{a}\right)$. To see the problem, we first consider the following failed parameterization motivated by our successful parameterization of marginal structural mean models. Let $q\left(y, \overline{\ell}_m, \overline{a}, a_m^*\right)$ be defined by

$$pr\left[q(Y_{\overline{a}}, \overline{\ell}_m, \overline{a}, a_m^*) < y \mid \overline{\ell}_m, \overline{a}_m\right] = pr\left[Y_{\overline{a}} < y \mid \overline{\ell}_m, \overline{a}_{m-1}, a_m^*\right] . \tag{8.98}$$

Let $\nu^*\left(y, \overline{\ell}_m, \overline{a}\right)$ be defined by

$$pr\left[\nu^*(Y_{\overline{a}}, \overline{\ell}_m, \overline{a}) > y \mid \overline{\ell}_m, \overline{a}_{m-1}\right] = pr\left[Y_{\overline{a}} > y \mid \overline{\ell}_{m-1}, \overline{a}_{m-1}, \ell_m \backslash v_0^* = 0\right], \tag{8.99}$$

where $\ell_m \backslash v_0^* = \ell_m$ for $m \neq 0$. Now consider sets of functions.

$$\left\{q\left(y, \overline{\ell}_m, \overline{a}, a_m^*; \mu_1\right); q\left(y, \overline{\ell}_m, \overline{a}, a_m; \mu_1\right) = y \text{ and } \mu_1 \in \boldsymbol{\mu}_1\right\} \tag{8.100a}$$

$$\left\{\nu^*(y, \overline{\ell}_m, \overline{a}; \mu_2); \nu^*(y, \overline{\ell}_{m-1}, \ell_m \Big\backslash v_0^* = 0, \overline{a}; \mu_2) = y \text{ and } \mu_2 \in \boldsymbol{\mu}_2\right\} \tag{8.100b}$$

and densities

$$\left\{f\left(\ell_k \mid \overline{\ell}_{k-1}, \overline{a}_{k-1}; \gamma\right); \gamma \in \boldsymbol{\gamma}, k = 0, \ldots, K\right\} . \tag{8.100c}$$

Then, using (8.100), (8.44), and (8.90), the likelihood function is, with $\rho = (\psi, \gamma, \eta, \mu_1, \mu_2)$,

$$\begin{aligned} f(O;\rho) = &\left\{\partial M_0^*(\rho)/\partial Y\right\} f\left[M_0^*(\rho) \mid V_0^*; \psi\right] \\ &\times \prod_{m=0}^{K} f\left[A_k \mid \overline{L}_k, \overline{A}_{k-1}; \eta\right] \prod_{m=0}^{K} f\left[L_k \mid \overline{L}_{k-1}, \overline{A}_{k-1}; \gamma\right] . \end{aligned} \tag{8.101}$$

As indicated in (8.101), there will almost always be at most one law of O and at most one random variable $M_0^*(\rho)$ associated with each choice of ρ. In particular, this will be so if $\nu^*\left(y, \overline{\ell}_m, \overline{a}; \mu_2\right)$ and $q\left(y, \overline{\ell}_m, \overline{a}, a_m; \mu_1\right)$

tend to ∞ as $y \to \infty$. However, this parameterization is not variation-independent. That is, there are parameters $\rho \in \boldsymbol{\rho} = \boldsymbol{\psi} \times \boldsymbol{\gamma} \times \boldsymbol{\eta} \times \boldsymbol{\mu}_1 \times \boldsymbol{\mu}_2$ that fail to correspond to any joint distribution. The problem is that in order to compute $m\left(y, \bar{\ell}_m, \bar{a}\right)$ and $m^*\left(y, \bar{\ell}_m, \bar{a}\right)$ in terms of the components of ρ, we need to solve integral equations that may not admit solutions. The basic problem is that the following conjecture is false.

CONJECTURE 8.1. *Given any function $q(y,x)$ satisfying $q(y,0) = y$ and increasing in y, a continuous distribution function $F(y)$, and a distribution $G(x)$ for X, there exists a joint law for (X,Y) with Y marginally distributed F, X with marginal G, and with $q(y,x)$ the quantile-quantile function linking F_x with F_0, i.e., $q(y,x) = F_x^{-1}\{F_0(y)\}$ where $F_x(y)$ is the law of Y given $X = x$ and 0 is in the support of X.*

Robins (1997b) had earlier proposed an alternative parameterization that he claimed was variation-independent. Unfortunately, this claim was incorrect. Specifically, on pg. 114 of Robins (1997b), there is suggested a parameterization, in the notation of that paper, in terms of $b(y,\bar{a})$ and $\nu^*(y,\bar{\ell}_m,\bar{a})$ which is used to obtain the quantile-quantile function $\nu(y,\bar{\ell}_m,\bar{a})$. This quantile-quantile function is analogous to our $m^*(y,\bar{\ell}_m,\bar{a})$. The difficulty is that Robins assumes that $\nu\left(y,\bar{\ell}_m,\bar{a}\right)$ so obtained is increasing in y and thus a valid quantile-quantile function. However, this is not proved and is in fact false in general.

Even though the above parameterization in terms of ρ is not variation-independent, it is nonetheless true that this submodel with q known [i.e., the set μ_1 in (8.100a) having but a single member] and the other components of ρ completely unrestricted is a non-parametric just-identified model for the law of F_O (even though certain values of ρ allowed by the model do not correspond to any joint distribution, which can create difficulties when fitting a fully parametric submodel by the method of maximum likelihood).

However, we can obtain a variation independent parametrization based on the following theorem.

THEOREM 8.16. *Given any non-negative function $q(y,x)$ satisfying $q(y,0) = 1$, a continuous distribution function $F(y)$, and a distribution $G(x)$ for X, there exists a joint law for (X,Y) with Y marginally distributed F, X with marginal G, and with $q(y,x)$ the relative risk (i.e., hazard ratio) function linking F_x with F_0, i.e., $q(y,x) = \lambda_x(y)/\lambda_0(y)$ where $\lambda_x(y)$ is the hazard of Y given $X = x$ and 0 is in the support of X.*

Proof. Define $\lambda_0(y)$ to be the unique solution to the Volterra-like integral equation

$$\lambda_0(y) = f(y) \Big/ \int \left\{ \exp\left[-\int_{-\infty}^{y} q(u,x)\lambda_0(u)\,du \right] \right\} q(y,x)\,dG(x) \ .$$

□

Then the joint law determined by $G(x)$, $q(y,x)$ and $\lambda_0(y)$ is such a joint law since upon multiplying both sides of the last display by the

negative of the denominator and integrating w.r.t. to y we obtain, as required,

$$1 - F(y) = \int \left\{ \exp\left[- \int_{-\infty}^{y} q(u, x) \lambda_0 (u)\, du \right] \right\} dG(x) \ .$$

It follows that we could obtain a variation independent parametrization by redefining

$$q\left(y, \bar{\ell}_m, \bar{a}, a_m^*\right) = \lambda_{Y_{\bar{a}}}\left[y \mid \bar{\ell}_m, \bar{a}_{m-1}, a_m^*\right] \Big/ \lambda_{Y_{\bar{a}}}\left[y \mid \bar{\ell}_m, \bar{a}_m\right]$$

and

$$\nu^*\left(y, \bar{\ell}_m, \bar{a}\right) = \lambda_{Y_{\bar{a}}}\left[y \mid \bar{\ell}_m, \bar{a}_{m-1}\right] \Big/ \lambda_{Y_{\bar{a}}}\left[y \mid \bar{\ell}_{m-1}, \bar{a}_{m-1}, \ell_m \Big\backslash v_0^* = 0\right]$$

where $\lambda_{Y_{\bar{a}}}[y \mid \cdot]$ is the hazard of the random variable $Y_{\bar{a}}$. Then (8.100a) and (8.100b) need only be modified by replacing "$= y$" by "$= 1$." Note this parametrization contains the implicit assumption that the above hazard ratios are finite, which thus restricts the magnitude of selection bias that can be represented, since, for any choice of the rate ratio function $q\left(y, \bar{\ell}_m, \bar{a}, a_m^*\right)$, the support of $Y_{\bar{a}}$ among subjects with history $\bar{\ell}_m, \bar{a}_{m-1}$ is contained within that for subjects with history $\bar{\ell}_{m-1}, \bar{a}_{m-1}, \ell_m \backslash v_0^* = 0$. The previous parametrization where $q\left(y, \bar{\ell}_m, \bar{a}, a_m^*\right)$ is a quantile-quantile function does not imply a similar restriction. Furthermore in the model in which the rate ratio selection bias function $q\left(y, \bar{\ell}_m, \bar{a}, a_m^*\right)$ is specified and regarded as known, we could, in principle, reject the hypothesis that $\nu^*\left(y, \bar{\ell}_m, \bar{a}\right)$ is everywhere finite, as the hazards $\lambda_{Y_{\bar{a}}}\left[y \mid \bar{\ell}_m, \bar{a}_{m-1}\right]$ are identified from the law of the observed data, and the hazard ratio defining $\nu^*\left(y, \bar{\ell}_m, \bar{a}\right)$ may be infinite.

8.7.2. Semiparametric inference in marginal structural distribution models. Robins (1998b, Appendix 3) provides semiparametric estimators of the finite dimensional parameter β of a marginal structural distribution model in which

(8.102) $q\left(y, \bar{\ell}_m, \bar{a}, a_m^*\right)$ *of Eq.(8.98) is known*

(8.103) $\psi = (\beta, \theta)$ *in (8.90) is composed of a finite-dimensional parameter* of interestβ *and an infinite-dimensional nuisance parameter* θ

(8.104) η *in (8.44a) is a finite-dimensional vector.*

Specifically, Robins (1998b) gives a semiparametric estimation algorithm that provides regular, asymptotically linear estimators of β. However, in

contrast to semiparametric estimation of marginal structural mean models, because the parameterization ρ described in the previous subsection is not variation-independent, it is possible that there will exist no joint distribution compatible with the estimates $\widehat{\beta}$, $\widehat{\eta}$ and the specified selection bias function q. We do not know of a simple way to check for mutual compatibility of resulting estimates and selection bias functions. Unfortunately, the variation independent parametrization described above in which the selection bias function $q\left(y, \overline{\ell}_m, \overline{a}, a_m^*\right)$ is a known hazard ratio function rather than the quantile-quantile function of Eq. (8.98) does not allow for estimation of the parameter β of this model without "high dimensional nonparametric smoothing," which is not feasible in the moderate size samples occurring in practice.

9. A General non-parametric identified (NPI) model with non-monotone non-ignorable missing data. In this section, we return to missing data models. The missing data models discussed in Secs. 2–5 assumed a monotone missing data pattern. In this section, we consider non-monotone missing data patterns. For the most part, we concentrate on settings in which there is a positive probability for each subject of observing a complete observation, without any missing data.

Robins (1997a) and Gill and Robins (1997) proposed a class of NPI models for non-monotone non-ignorable missing data called the Group Permutation Missingness (GPM) models by Robins (1997a) and Sequential Coarsening at Random models by Gill and Robins (1997). The GPM models are representable by a sequence of nested coarsening at random (CAR) models. Interestingly, a sequence of CAR models is, in general, not itself CAR. As noted by Robins (1997a), the class of GPM models has a serious drawback that prevents it from serving as an all-purpose class of missing data models with which to model non-monotone non-ignorable missing data. Specifically, GPM models do not allow for the probability that a particular variable is missing to depend on the value of that variable, although this probability can depend on the values of other missing variables. The NPI models described in Section 9.1 overcome the above deficiency of the GPM models. However, as we will see, new difficulties arise. These new difficulties are largely solved by introducing a new class of NPI models — The Selection Bias Permutation Missingness Models (PM) of Sec. 9.2.

9.1. A class of NPI models. The results in this subsection are based on joint unpublished work with Richard Gill.

Gill et al. (1997) prove in their Theorem 1 that CAR models are NPI. Our new models are based on the following generalization of their Theorem.

Let X be a discrete random variable with sample space E. We observe χ, a random subset of E. Let A denote a realization of χ. Consider the model

$$
\begin{aligned}
pr[\chi = A \mid X = x] &= \pi_A q_A(x) \ , \quad x \in A, \\
pr[\chi = A \mid X = x] &= 0 \ , \quad x \notin A,
\end{aligned} \tag{9.1}
$$

where $q_A(x)$ is a known function of x and A, π_A is a completely unknown function of A, and the sample space E is also unknown. Thus, for each A, $q_A(x)$ is a "known" selection bias function that we vary in a sensitivity analysis. Note if $q_A(x) = 1$ for all (A, x) (or more generally $q_A(x)$ does not depend on x), then model (9.1) is a NPI CAR model. Let $f_A = pr(\chi = A)$ denote the law of the observed data χ.

THEOREM 9.1. *Given model* (9.1) *and* f_A, *there exists* π_A *and a distribution* p_x *on some discrete sample space* E *such that* $f_A = p_A \pi_A$ *for all* A *in the power set of* E, *where* $p_A = \sum_{x \in A} q_A(x) p_x$, *for each* $x \in E$

$$
\sum_{\{A; A \ni x\}} \pi_A q_A(x) = 1 \ , \tag{9.2}
$$

$\sum_{x \in E} p_x = 1$, $p_x > 0$, $\pi_A \geq 0$. *Furthermore,* π_A *and* p_A *are unique if* $f_A > 0$.

REMARK 9.1. As discussed below, there are two differences between Theorem 9.1 and Theorem 1 of Gill et al. (1997). The first and most obvious is that $q_A(x)$ may be a function of x for $x \in A$, in which case model (9.1) is non-ignorable (i.e., non-CAR). Secondly, the sample space E is not given beforehand. Rather it is determined by the requirement that $p_x > 0$. In contrast, in Theorem 1 of Gill et al., E was given beforehand and $p_x > 0$ was replaced by $p_x \geq 0$. As we will see below, Theorem (9.1) is false if E is given beforehand.

The proof of Theorem 9.1 follows exactly the proof of Theorem 1 in Gill et al. (1997). We obtain p_x and E by maximizing $\sum_A f_A \log p_A$ subject to the constraint that $\sum_{x \in E} p_x = 1, p_x > 0$ and then, defining $\pi_A = f_A / p_A$ for $p_A \neq 0$. Theorem 9.1 implies that model (9.1) is a non-parametric model for the observed data χ. Furthermore, if, for all $x \in E$, $f_{\{x\}} \neq 0$ (i.e., there is a non-zero probability of observing any singleton), then the distribution p_x is identified. As we will show below, it is possible that p_x is identified even when there is not a positive probability of observing any singleton. An example will clarify the theorem.

EXAMPLE 1. Let $X = (X_0, X_1)$, where X_0 and X_1 are dichotomous $(0, 1)$ variables. Suppose there are four events A with positive probability. Specifically, observe X_0 only and it takes value 1; observe X_0 only and it is 0; observe X_1 only and it is 1; observe X_1 and it is 0. This example models a study where X_0 and X_1 are counterfactual dichotomous outcomes corresponding to a control treatment 0 and an active treatment 1,respectively, so you always observe one but not both of X_0 and X_1. If treatment has been randomly assigned, we have CAR. Otherwise, we have an observational study with non-ignorable missingness (i.e., confounding by unmeasured factors). The four possible events can be characterized

as $X_0 = 1, X_0 = 0, X_1 = 1, X_1 = 0$, which we denote by $A = 1, 2, 3, 4$, respectively.

TABLE 1

		X			
		$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
A	$X_0 = 1$	0	0	$\pi_1 q_1 (1,0)$	$\pi_1 q_1 (1,1)$
	$X_0 = 0$	$\pi_2 q_2 (0,0)$	$\pi_2 q_2 (0,1)$	0	0
	$X_1 = 1$	0	$\pi_3 q_3 (0,1)$	0	$\pi_3 q_3 (1,1)$
	$X_1 = 0$	$\pi_4 q_4 (0,0)$	0	$\pi_4 q_4 (1,0)$	0
	E	1	1	1	1

We assume the law f_A is given and model (9.1) holds, so $q_A(x)$ and thus, by Theorem 9.1, the π_A are known for $A \in \{1, 2, 3, 4\}$. Now consider Table 1. The four events with non-zero probability are given in column 1. The entries in the first four rows are $pr[\chi = A \mid X = (x_0, x_1)] \equiv \pi_A q_A (x_0, x_1)$. The fifth row is all 1's. Let M denote the 5×4 matrix given by the interior of the table. Then, by the laws of probability,

$$M (p_{00}, p_{01}, p_{10}, p_{11})' = (f_1, f_2, f_3, f_4, 1)', \tag{9.3}$$

where, e.g., p_{00} is $p_{x_0=0, x_1=0}$ and $f_1 = f_{A=1}$. Under the CAR model in which treatment was randomly assigned [i.e., $g_A(x) \equiv 1$], M is of rank 3 so that the four unknowns $p = (p_{00}, p_{01}, p_{10}, p_{11})'$ are not identified by (9.3). This corresponds to the fact that it is not possible to identify the joint distribution of (X_0, X_1) in a randomized trial, although one can identify $pr(X_1 = 0) = p_{00} + p_{10}$ and $pr[X_0 = 0] = p_{01} + p_{00}$.

However, when the non-zero entries in some row are not equal (i.e., $q_A(x)$ depends on x for some $A \in \{1, 2, 3, 4\}$), M can be of rank 4 in which case the joint law p_x of $X = (X_0, X_1)$ will be identified, even though the model (9.1) is non-parametric for χ, no restrictions are imposed on the joint distribution of (X_0, X_1), and X_0 and X_1 are never simultaneously observed.

As odd as this may seem, things may become even odder. For example, consider model (9.1) with $q_3(1,1) = b \neq 1$ and $q_A(x_0, x_1) = 1$ for all other A and x. Now, as conditional probabilities given X, the first four rows in each of the columns of M must sum to the value 1 in the fifth row, (i.e., Eq. (9.2) must hold). This implies that

$$\pi_2 + \pi_4 = 1, \quad \pi_2 + \pi_3 = 1, \quad \pi_1 + \pi_4 = 1, \quad \pi_1 + b\pi_3 = 1\ . \tag{9.4}$$

If, as we have assumed, $b \neq 1$, we conclude from (9.4) that $\pi_3 = \pi_4 = 0$ whatever be the law f_A. However, from the third and fourth rows of M, we see that $\pi_3 = \pi_4 = 0$ implies that the probability that we observe X_1

must be zero (i.e., $f_3 + f_4 = 0$), which may be contradicted by the given law f_A. Thus it appears that model (9.1) must be misspecified since the choice of $q_A(x)$ is incompatible with f_A. However, according to Theorem 9.1, model (9.1) is nonparametric. The resolution of the paradox is that the sample space E for which Theorem 9.1 is true will exclude at least one of the four possible values of $x = (x_0, x_1)$, i.e., for at least one of these values of x, $p_x = 0$. In that case, at most three of the four equations in (9.4) must hold which will not restrict the possible f_A, since Theorem 1 does not require Eq. (9.2) to hold for values of x that are not in the sample space E (i.e., for values of x for which $p_x = 0$). Note that Theorem 1 of Gill et al. proves that in the case in which model (9.1) is CAR, it is never necessary to delete any points from an *a priori* sample space E in order that Eq. (9.2) hold.

As odd as the above results may seem, model (9.1) has two even greater problems that will probably make it unsuitable for use in sensitivity analysis. First, it is quite unclear what the substantive meaning of $q_A(x)$ is in a given problem, thus making it hard to choose plausible functions $q_A(x)$ for sensitivity analysis. Second, suppose there is a positive probability of observing the full data X and our analytic strategy is to first model the non-response mechanism and then. to estimate the law of X by inverse probability weighting. If X is high dimensional, it is difficult to know how to do dimension reduction in model (9.1). Specifically, if we specify a model $\pi_A(\alpha)$ (depending on a finite dimensional parameter α) for the unknown π_A and then estimate α by maximizing $\sum_A f_A \log \pi_A(\alpha)$, it is unlikely that (9.2) will hold with our estimate replacing π_A. Indeed, there may be no parameter value α for which (9.2) holds with π_A replaced by $\pi_A(\alpha)$, which would imply that we could conclude that our model $\pi_A(\alpha)$ is misspecified even before seeing the data χ. This same difficulty arises even with CAR models.

9.2. Selection bias permutation missingness models. As in Robins and Gill (1997) and Robins (1997a), we assume data $L = (L_0, \ldots, L_K)'$ and let R_k be the indicator that variable L_k is observed. The observed data is $(R, L_{(R)})$ where $L_{(r)}$ are the observed components of L when $R = r$. We assume that there is a positive probability of complete observations, i.e., $pr[R = \mathbf{1} \mid L] > 0$ with probability 1, where $\mathbf{1}$ is the vector of 1's. To be able to use the notation of Robins and Gill (1997), we assume, as they did, that $L = (Y, X')'$ where $Y = L_0$ is an always observed variable and X may have one or more components missing. We shall obtain a separate NPI model for each of the $K!$ permutations of the variables $(X_1, \ldots, X_K)$. Given a permutation, let $(X^{(1)}, \ldots, X^{(K)})$ denote the first to last variables in our permutation. Let R^k denote the indicator of whether variable $X^{(k)}$ is observed. Define $W_k = \{R^{k+1}, \ldots, R^k, X^{(1)}, \ldots, X^{(k-1)}, R^{(k+1)}X^{(k+1)}, \ldots, R^K X^{(K)}, Y\}$. Then a NPI selection odds PM model specifies

$$\text{(9.5)} \qquad \textit{logit pr}\left[R^k = 1 \mid W_k, X^{(k)}\right] = h_k\left(W_k\right) + q_k\left(X^{(k)}, W_k\right)$$

with

$$\text{(9.6)} \qquad q_k\left(X^{(k)}, W_k\right) \textit{ known}$$

and

$$\text{(9.7)} \qquad h_k\left(W_k\right) \textit{ unrestricted .}$$

Robins (1997a) shows that this model is a NPI model in the special case in which q_k is identically zero. In this case, he shows also how to carry out estimation in the restricted model characterized by (9.5), (9.6), and

$$\text{(9.8)} \qquad h_k\left(W_k\right) = h_k\left(W_k; \gamma_k^*\right), \quad k = 1, \ldots, K$$

where $h_k\left(W_k; \gamma_k\right)$ is a known function and γ_k^* is an unknown finite dimensional vector to be estimated. Specifically, Robins (1997a) shows how to estimate γ_k^* by sequential inverse-probability-weighted-estimators and then estimate the functionals of the joint distribution of the full data (Y, X) by inverse-probability-weighting as well.

We can use this same strategy for selection odds PM models under the additional restriction (9.8). Specifically, note that for $k = 1$, we have an ordinary selection odds model, since W_1 is completely observed. Thus, we can consistently estimate γ_1^* using the inverse-probability-of-censoring-weighted methods of Rotnitzky, Robins, and Scharfstein (1998) and Secs. 2–4. Now consider $k = 2$. This would also be a selection odds model of the type studied in Secs. 2–4, except that the component $X^{(1)}$ of W_2 is not completely observed. However, we can restrict attention to this subset of the population in which $X^{(1)}$ is completely observed by reweighting by our estimate of $pr\left[R^1 = 1 \mid W_1, X^{(1)}\right]$. We can then proceed this way recursively until all the γ_k^*'s have been estimated. The strategy is exactly that described on page 27 of Robins (1997a) for the special case in which q_k was zero.

10. A Non-ignorable Generalization of RMM Models. Robins and Gill (1997) when faced with the difficulty with CAR models described in the last paragraph of Sec. 9.1, proposed to solve it by introducing the class of randomized monotone missingness (RMM) models, a subclass of the class of CAR models. In this section, we introduce a non-CAR generalization of Markov RMM models — The non-ignorable selection odds Markov RMM models. We first review the definition of a Markov RMM model.

Consider Figure 3, taken from Robins and Gill (1997). In Figure 3, at stage m, $m = 1, 2, \ldots, M+1$, there are $\binom{M}{m-1} = M!/\{m-1\}!\left[M-\right.$

$(m-1)]!$ groups of $m-1$ variables $X^{mk}, k = 1, \ldots, \binom{M}{m-1}$. For example, at stage $m = 3$, we have $3!/(2!\,1!) = 3$ groups of $m - 1 = 2$ variables. Each group $X^{mk}, m \leq M$, is connected by arrows to the $M - (m-1)$ groups $\{X_j, X^{mk}\}$ at stage $(m+1)$ with $X_j \notin X^{mk}$. For example, if $X^{21} = \{X_1\}$, X^{21} is connected to the $3 - (2-1) = 2$ groups $\{X_1, X_2\}$ and $\{X_1, X_3\}$. The probabilities $p_j\left(X^{mk}, Y\right)$ on Figure 3 are the conditional probabilities that the variable $X_j, X_j \notin X^{mk}$, will be observed in the next stage conditional on the observed values of Y and of the variables X^{mk} that have been observed through stage m. (The dependence of these probabilities on the always-observed variable Y is suppressed in the figure.) $p_-\left(X^{mk}, Y\right) \equiv 1 - \sum_{\{j; X_j \notin X^{mk}\}} p_j\left(X^{mk}, Y\right)$ is the conditional probability of quitting without proceeding to the next stage. A parametric Markov RMM models specifies a parametric form for the unknown $p_j\left(X^{mk}, Y\right)$.

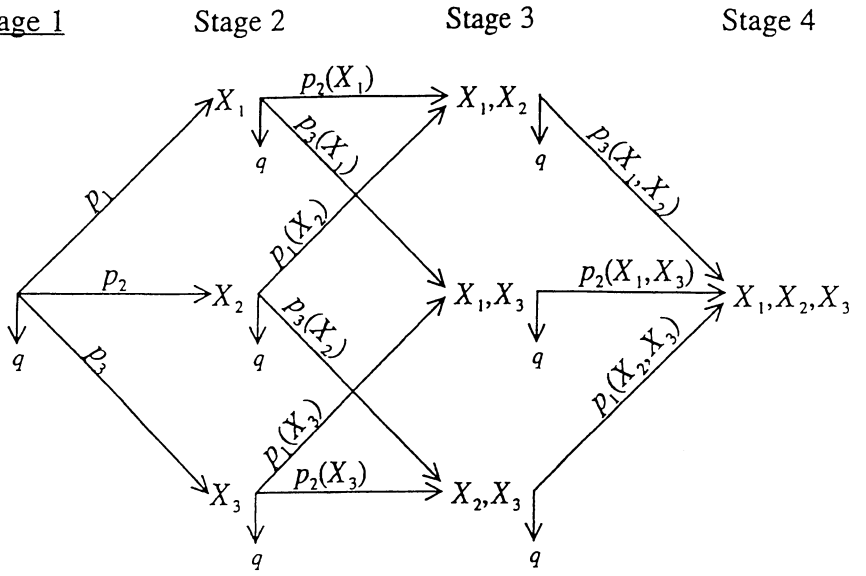

FIG. 3.

We can generalize this Markov RMM to a non-ignorable selection odds model. We now redefine $p_j\left(X^{mk}, Y\right)$ to be $p_j\left(X^{mk}, Y, \underline{X}^{mk}\right)$ where $\underline{X}^{mk}$ are the components of X other than X^{mk}. Then an unrestricted (possibly) non-ignorable selection odds Markov RMM model assumes

$$p_j\left(X^{mk}, Y, \underline{X}^{mk}\right) = \phi_{jmk} \Big/ \left\{1 + \sum_{\{j; X_j \notin X^{mk}\}} \phi_{jmk}\right\} \tag{10.1}$$

where

$$\phi_{jmk} = \exp\left[h_{jmk}\left(X^{mk}, Y\right) + q_{jmk}\left(X, Y\right)\right] \tag{10.2}$$

and the q_{jmk} are known functions and the functions h_{jmk} are completely unknown. Robins and Gill (1997) showed in the special case where $q_{jmk}(X, Y)$ does not depend on $\underline{X}^{mk}$ (so the model is CAR) an unrestricted RMM model is not a nonparametric model for the law of the observed data.

In practice, to overcome the curse of dimensionality, we specify that

$$h_{jmk}\left(X^{mk}, Y\right) = h_{jmk}\left(X^{mk}, Y; \gamma\right) \tag{10.3}$$

where γ is an unknown finite-dimensional parameter and $h_{jmk}(\cdot, \cdot; \gamma)$ is a known function. As in Robins and Gill (1997), our goal is to use our non-ignorable selection odds Markov RMM models to first estimate γ and then, under the assumption that

$$pr\left[R = \mathbf{1} \mid L\right] > \sigma > 0 \ w.p.1, \tag{10.4}$$

estimate the distribution of X by inverse probability weighting.

To estimate the law of L by inverse probability weighting, it is necessary that $pr[R = \mathbf{1} \mid L]$ be identified where $\mathbf{1}$ is a vector of ones. At present, it is an open question whether the unrestricted non-ignorable selection odds Markov RMM given by (10.1) and (10.2) is sufficient to identify $pr[R = \mathbf{1} \mid L]$ under assumption (10.4). Note that even with X and Y discrete, the functions h_{jmk} will not be identified. However, we conjecture that the probability $pr[R = r \mid L]$ will be identified under model (10.1)–(10.2).

We now discuss how one might estimate parameters γ of (10.3) of our dimension-reduced model. Note that (10.1)–(10.3) imply that $pr[R = r \mid L]$ is a known function $pr[R = r \mid L; \gamma]$ of L and γ. Thus we consider estimating γ by the

$$\sum_i B_i(\phi, \gamma) = 0 \tag{10.5}$$

where $B(\phi, \gamma) = \sum_r \left(I(R = r) - I(R = \mathbf{1})\, pr[R = r \mid L; \gamma] / pr\left[R = \mathbf{1} \mid L; \gamma\right]\right)\phi_r\left(L_{(r)}\right)$ where the $\phi_{(r)} \mid \cdot$ are functions chosen by the data analyst. In general, γ may not be identified, in which case Eq. (10.5) may have many solutions. However, we conjecture under regularity conditions, all such solutions $\widehat{\gamma}$ imply the same value for $pr[R = r \mid L; \widehat{\gamma}]$. When a dimension K of L is large, solving (10.5) will be computationally intractable because $B(\phi, \gamma)$ is a sum over 2^Kterms . It will be of interest to determine if computationally tractable simulation methods exist for approximately solving (10.5) can be derived.

11. Sensitivity Analysis and Bayesian Inference.

11.1. A parametric missing data example. If a decision is required, one may wish to consider a Bayesian analysis in place of a sensitivity analysis. We shall consider the following simple setting. We return to the missing data setting of Sec. 2. Let $K = 1$, $L_0 = Z$, $L_1 = W$, $L_2 = Y$, $L = \overline{L}_2 = (Z, W, Y)$. We assume the censoring-time C is either at times 1 or 2 with probability 1. The observed data are

$$O = (Z, W, \Delta, \Delta Y)$$

where $\Delta = 1 \Longleftrightarrow C = 2$. We consider the model

$$pr\,[\Delta = 1 \mid L] = h\,(Z, W) + q\,(L) \tag{11.1}$$

where

$$h\,(\overline{\ell}_1) = h\,(z, w) \ \textit{is unknown} \tag{11.2}$$

and

$$q\,(\ell) \ \textit{is known.} \tag{11.3}$$

We know from Theorem 2.2 that the semiparametric model **a** determined by the restrictions (11.1)–(11.3) is a non-parametric model for the observed data O and that $F_{\Delta,L}$ is identified.

The setting we are thinking of here is a randomized clinical trial where *(i)* $Z \in (0, 1)$ is the treatment arm randomization indicator, *(ii)* W is an always observed post-randomization variable that is discrete with d points of support, and *(iii)* Y is an outcome of interest which is not observed for some subjects who have dropped out. Thus, our parameter of interest is

$$\beta \equiv E\,(Y \mid Z = 1) - E\,(Y \mid Z = 0)\ ,$$

the effect of treatment on the outcome Y.

Note, by assuming W is discrete, we are considering the case where we are not afflicted by the curse of dimensionality.

To perform a Bayesian analysis, rather than treating $q \equiv q\,(\ell)$ as known and varying it in a sensitivity analysis, we shall put a prior $\pi\,(q)$ on the function q in (11.1). Since $L = (W, Z, Y)$ has $4d$ points of support, a particular function q can be identified with a point in $\mathcal{R}^{4d}$. Similarly, $h = h\,(Z, W)$ can be identified with a point in $\mathcal{R}^{2d}$, so when we only impose (11.1), q and h can be viewed as unknown parameters in $\mathcal{R}^{4d}$ and $\mathcal{R}^{2d}$. Thus, our prior $\pi\,(q)$ on q is simply a probability distribution on $\mathcal{R}^{4d}$.

We now consider approximate Bayesian inference on β. Let $\beta\,(q)$ be the value of β defined by the law F_O of O and the known function q in the semiparametric model **a** characterized by (11.1)–(11.3). Let $\widehat{\beta}\,(q)$ be a semiparametric efficient estimator of β under model **a**. To compute $\widehat{\beta}\,(q)$, one could use the discrete version of the methods described in Sec. 4, or

one could estimate $\widehat{\beta}(q)$ by non-parametric maximum likelihood. We shall suppose the sample size n is sufficiently large that, given q is known, the sampling distribution of $\widehat{\beta}(q)$ is approximately normal with mean $\beta(q)$ and standard error that can be consistently estimated by $\widehat{\sigma}(q)$, say, using the methods described in Sec. 4. Then *(i)*, since, given q, $\beta(q)$ is identified and *(ii)* for large n, the data should dominate the prior, any reasonable robust Bayesian procedure should by the Bernstein-von Mises theorem result in inferences such that, given both q known and the data [i.e., $\{O_i; i = 1, \ldots, n\}$], the posterior distribution of β will be approximately normal with mean $\widehat{\beta}(q)$ and variance $\widehat{\sigma}^2(q)$. Indeed, any prior for which the above approximation of the posterior is inadequate will be suspect, as the prior will have dominated the data.

Hence the following algorithm gives an asymptotic approximation $\widetilde{\pi}(\beta \mid data)$ to the posterior density $\pi(\beta \mid data)$ of β.

Algorithm: For $j = 1, \ldots, J$, *(i)* draw q_j from $\pi(q \mid data)$, where $\pi(q \mid data)$ is the posterior density of q given the data $\{O_i; i = 1, \ldots, n\}$, and *(ii)* compute $\widetilde{\pi}(\beta \mid data) = J^{-1} \sum_{j=1}^{J} \phi\left(\beta; \widehat{\beta}(q_j), \widehat{\sigma}^2(q_j)\right)$ where $\phi\left(\beta; \mu, \sigma^2\right)$ is a $N\left(\mu, \sigma^2\right)$ density evaluated at β.

Thus it only remains to determine how to draw from the posterior distribution of q given the data. Since q is not identified in the model characterized by (11.1) with both h and q unknown, the posterior distribution $\pi(q \mid data)$ of q equals the prior distribution if and only if *a priori* q is independent of the distribution F_O of the observed data [i.e., $\pi(F_O, q) = \pi(F_O)\pi(q)$]. Given such independence, we can simply draw q from its prior $\pi(q)$. However, we now argue that it may be more natural and substantively plausible not to assume that F_O and q are *a priori* independent. Note that the joint distribution $F_{\Delta,L}$ of (Δ, L) can be written $F_{\Delta,L} = (h, \beta, \gamma, \theta, q) \equiv (\eta, q)$ where γ parameterizes the distribution $f(W \mid Y, Z)$ and can be represented as a point in $\mathcal{R}^{4(d-1)}$, $\theta = E[Y \mid Z = 0]$, and $\eta \equiv (h, \beta, \gamma, \theta)$. Now as a marginal law, the law F_O of O is a function, say s, of $F_{\Delta,L}$ i.e., $F_O = s(\eta, q)$. In general, $s(\cdot, \cdot)$ will be a smooth differentiable map between two Euclidean spaces. By Theorem 2.2, we know that, for fixed q, $s(\eta, q)$ is one-to-one in η, since Theorem 2.2 states that q and F_O determine $F_{\Delta,L} = (\eta, q)$. That is, $\eta = s^{-1}(F_O, q)$. Note that if, we for example, assume that η is *a priori* independent of q, i.e., $\pi(\eta, q) = \pi(\eta)\pi(q)$, then $F_O = s(\eta, q)$ and q will not be *a priori* independent and $\pi(q \mid data) \neq \pi(q)$. We believe it is more natural to specify a prior $\pi(\eta, q)$ for the distribution of $F_{\Delta,L}$ than to directly specify a prior distributions for F_O. Under such prior specification, we now develop an approach to sampling from $\pi[q \mid data]$. Given a prior $\pi(\eta, q)$, let

$$\omega(q) = \pi(q \mid data)/\pi(q)$$

be the posterior-prior importance weights. Let $\widehat{F}_O$ be the empirical distribution of the $O_i, i = 1, \ldots, n$.

THEOREM 11.1. *Suppose that* $\pi(\eta, q)$ *is continuous in* (η, q). *Then* $\widehat{\omega}(q) = c\omega(q) + o_p(1)$, *where*

$$\widehat{\omega}(q) = \pi_{\eta|q}[\widehat{\eta}(q) \mid q] \mid \partial s(\eta, q)/\partial \eta \mid^{-1}_{\eta=\widehat{\eta}(q)} \tag{11.4}$$

where c *is an unknown normalizing constant,* $\widehat{\eta}(q) = s^{-1}\left(\widehat{F}_O, q\right)$ *is the non-parametric maximum likelihood estimator of* η *when* q *is known,* $|A|$ *is the Jacobian determinant of* A, $\pi_{\eta|q}$ *is the conditional density of* η *given* q, *and* $o_p(1)$ *is a random variable converging to zero under* F_O.

It follows from Theorem 11.1 that we can consistently estimate $\omega(q)$ up to a normalizing constant. Thus, we can modify our previous algorithm and use the following approximation $\pi^*(\beta \mid data)$ to the posterior $\pi(\beta \mid data)$ suggested by L. Wasserman.

Modified algorithm: For $j = 1, \ldots, J$, *(i)* draw q_j from $\pi(q)$, and *(ii)* compute $\pi^*(\beta \mid data) = h(\beta) / \int h(\beta)\, d\beta$ where $h(\beta) = J^{-1} \sum_{j=1}^{J} \phi\left(\beta; \widehat{\beta}(q_j), \widehat{\sigma}^2(q_j)\right) \widehat{\omega}(q_j)$ and $\int h(\beta)\, d\beta$ can be evaluated numerically.

Note $\pi^*(q \mid data)$ will converge to $\pi(q \mid data)$ as $n \to \infty$ and $J \to \infty$.

In practice, we use our large sample approximation $\widehat{\omega}(q)$ in place of $\omega(q)$.

REMARK 11.1. In usual Bayesian inference, we cannot sample from $\pi(q \mid data)$ by sampling from $\pi(q)$ and then using importance weights $\omega(q)$, since $\pi(q \mid data)$ is a much more peaked distribution than $\pi(q)$ when the sample size is large. However, in our setting, since q is not identified in the model characterized by (11.1a) alone, $\pi(q \mid data)$ will not be a highly peak function of q even as $n \to \infty$. Indeed, we can empirically check whether it is reasonable to importance sample from the prior by plotting the distribution of the $\widehat{\omega}\left(q_j^*\right)$. If this distribution is not too highly variable nor too skew, our importance sampling approach is adequate.

Proof of Theorem 11.1. $\pi(q \mid data) = \int \pi(q \mid F_O)\pi(F_O \mid data)\, d\mu(F_O)$ $= \pi\left(q \mid \widehat{F}_O\right) + o_p(1)$ since *(i)* $q \coprod data \mid F_O$, and *(ii)* $\pi\left(F_O \mid data\right) = \delta_{\widehat{F}_O} + o_p(1)$ by the consistency of Bayes estimators, and, by assumption, $\pi\left(q \mid F_O\right)$ is smooth in F_O. Here $\delta_{\widehat{F}_O}$ is the distribution that puts all its mass $\widehat{F}_O$. Now $\pi\left(q \mid \widehat{F}_O\right) \propto \pi_{F_O|q}\left(\widehat{F}_O \mid q\right)\pi(q)$. So $\omega(q) = c\pi_{F_O|q}\left(\widehat{F}_O \mid q\right) + o_p(1)$. But, by the change of variables formula, $\pi_{F_O|q}\left(\widehat{F}_O \mid q\right) = \pi_{\eta|q}\left[s^{-1}\left(\widehat{F}_O, q\right) \mid q\right] \mid \partial s^{-1}\left(\widehat{F}_O, q\right)/\partial F_O \mid = \pi_{\eta|q}[\widehat{\eta}(q) \mid q] \mid \partial s(\eta, q)/\partial \eta \mid^{-1}_{\eta=\widehat{\eta}(q)}$. □

11.2. A parametric causal inference example.

11.2.1. Smooth priors. Consider the simple logistic SNMM of Sec. 8.4 with $K = 0$, $Y \equiv L_{K+1} = L_1$, $V \equiv L_0$, $A \equiv A_0$, with Y dichotomous, and V and A discrete with d_V and d_A points of support, respectively. Then (8.30) becomes

$$\gamma^*(v, a) = logit E(Y_a \mid a, v) - logit E(Y_0 \mid a, v) \ , \tag{11.5}$$

Eq. (8.11a) becomes

$$E(Y_0 \mid v, a) = expit[t(v) + q(v, a)] \tag{11.6}$$

where $q(v, 0) = 0$ and the logistic current treatment interaction function becomes

$$\begin{aligned} r(v, g) = &\{logitE[Y_g \mid v, A = g(v)] - logitE[Y_0 \mid v, A = g(v)]\} \\ &- \{logitE[Y_g \mid v, A \neq g(v)] - logitE[Y_0 \mid v, A \neq g(v)]\} . \end{aligned}$$

Note by the remarks following Eq. (8.29), the parameter space for $r(v, g)$ is precisely the set of all functions $r^*(v, a)$ satisfying $r^*(v, 0) = 0$. Hence, any function $r(v, g)$ can be represented as a point in $R^{d_V(d_A-1)}$. Further, since $K = 0$ and thus A is time-independent, Theorem (8.13) is true with $r(v, g)$ the logistic current treatment interaction function for all regimes g both dynamic and non-dynamic.

Now let $\eta = \left(F_V, F_{A|V}, \gamma^*, t\right)$ with $\gamma^* = \gamma^*(v, a)$ and $t = t(\ell)$. Let F_O be the joint distribution of (V, A, Y) which can be identified as a point in $\mathcal{R}^{2d_V d_A - 1}$. It follows from Theorems 8.5 and 8.13 that $F_O = s(\eta, q, r) = s(\eta, q)$ is a function of (η, q) alone. Further, given q, this function is invertible, i.e., $\eta = s^{-1}(F_O, q)$. Furthermore, $E(Y_g)$ is a deterministic function of (η, q, r).

Let $\beta = \beta(\eta, q, r)$ be a possible vector-valued functional of interest such as $E(Y_g)$ or a function of γ^*. We want to derive an approximation to the posterior distribution of β. For fixed F_O, define $\eta(q) = s^{-1}(F_O, q)$ and let $\widehat{\eta}(q) = s^{-1}\left(\widehat{F}_O, q\right)$ be the NPMLE of $\eta(q)$ where $\widehat{F}_O$ is the NPMLE of F_O. Arguing as in the last subsection, in large samples, the posterior distribution of η given the data and (q, r) will be approximately normal with mean $\widehat{\eta}(q)$ and variance $\widehat{\Sigma}(q)$ where $\widehat{\Sigma}(q)$ is a consistent estimator of the asymptotic variance of $\widehat{\eta}(q)$. Now let $\omega(q, r) = \pi(q, r \mid data)/\pi(q, r)$. Analogously to Theorem 11.1, we have

THEOREM 11.2. *Suppose the prior $\pi(\eta, q, r)$ is continuous. Then $\widehat{\omega}(q, r) = c\omega(q, r) + o_p(1)$ where c is a constant and*

$$\widehat{\omega}(q, r) = \pi_{\eta|q,r}[\widehat{\eta}(q) \mid q, r] \mid \partial s(\eta, q)/\partial \eta \mid^{-1}_{\eta=\widehat{\eta}(q)} . \tag{11.7}$$

By the delta method and the Bernstein-von Mises theorem, the posterior distribution of $\beta = \beta(\eta, q, r)$ given (q, r) will be asymptotically normal with mean $\widehat{\beta}(q, r) \equiv \beta(\widehat{\eta}(q), q, r)$ and variance $\widehat{\Sigma}_\beta(q, r)$ where $\widehat{\Sigma}_\beta(q, r) = \widehat{\tau}(q, r)\,\widehat{\Sigma}(q)\,\widehat{\tau}(q, r)$, $\widehat{\tau}(q, r) = \partial\beta(\eta, q, r)/\partial\eta \mid_{\eta=\widehat{\eta}(q)}$. Thus, we can consider the following approximation $\pi^*(\beta \mid data)$ for the posterior $\pi(\beta \mid data)$.

Algorithm: For $j = 1, \ldots, J$, *(i)* draw (q_j, r_j) from $\pi(q, r)$, and *(ii)* compute $\pi^*(\beta \mid data) = h(\beta) / \int h(\beta)\, d\beta$ where

$$h(\beta) = J^{-1} \sum_{j=1}^{J} \phi\left(\beta; \widehat{\beta}(q_j, r_j), \widehat{\Sigma}_\beta(q_j, r_j)\right) \widehat{\omega}(q_j, r_j) \ . \tag{11.8}$$

11.2.2. Allowing for a prior non-zero probability of non-causality. Heretofore, we have assumed our prior $\pi(\eta, q, r)$ was smooth. However, it is likely that one would wish to place positive prior mass on the causal null hypothesis that γ^* and r are zero, provided that treatment A was not already known to be a cause of Y. Therefore, define $M = 1$ to be the submodel in which γ^* and r are identically zero and let $M = 0$ denote the submodel in which neither is zero. For convenience, we shall assume a zero prior probability that only one of γ^* and r are precisely zero. Let $\pi(M = 1)$ be the prior probability that $M = 1$. We shall assume that the conditional prior $\pi(\eta, q, r \mid M = 0)$ is smooth except that we have deleted the point $\gamma^* = r = 0$. Let the parameter $\nu = \left(F_V, F_{A|V}, t, q\right)$ be the unknown parameter in model $M = 1$. We assume $\pi(\nu \mid M = 1)$ is a smooth prior. We now approximate the Bayes factor $\{\pi[M = 1 \mid data]/\pi[M = 0 \mid data]\}/\{\pi[M = 1]/\pi[M = 0] = f[data \mid M = 1]/f[data \mid M = 0]\}$. Since by Theorems 8.5 and 8.13, both models $M = 0$ and $M = 1$ are non-parametric models for the law F_O of the observed data, we know that, by Laplace's method,

$$\begin{aligned} &f[data \mid M = m] \\ &\quad \propto f\left[data \mid \widehat{F}_O\right] n^{-(2d_V d_A - 1)/2} \pi_{F_O|M=m}\left(\widehat{F}_O\right)\left(1 + O_p\left(n^{-1}\right)\right). \end{aligned} \tag{11.9}$$

Hence, $\dfrac{f[data \mid M = 1]}{f[data \mid M = 0]}$ is approximately

$$\frac{\pi_{F_O|M=1}\left(\widehat{F}_O\right)}{\pi_{F_O|M=0}\left(\widehat{F}_O\right)} = \frac{\pi_{\nu|M=1}(\widehat{\nu}) \left|\partial s^*(\widehat{\nu})/\partial \nu\right|^{-1}}{\int \pi[\widehat{\eta}(q), q, r \mid M = 0] \left|\partial s(\eta, q)/\partial \eta\right|^{-1}_{\eta=\widehat{\eta}(q)} dqdr}$$

where s^* is the one to one function mapping of ν into F_O under model $M = 1$ and $\widehat{\nu} = s^{*-1}\left(\widehat{F}_O\right)$.

Our goal remains to compute $f[\beta \mid data]$ as in Sec. 11.2.1. Given the above approximate formulas for the Bayes factors, we need approximate formulas for $f[\beta \mid M = 0, data]$ and $f[\beta \mid M = 1, data]$. Now $f[\beta \mid data, M = 0]$ is calculated just like $f[\beta \mid data]$ in Sec. 11.2.1, except now all conditioning events include $M = 0$. In model $M = 1$, the posterior distribution of β will be a point mass at zero whenever β is a causal contrast such as $E\left[Y_g - Y_g^*\right]$ since, in model $M = 1$, $\gamma^* = r = 0$. In model 1, if β is not a contrast [e.g., $\beta = E(Y_g)$], then $\beta \equiv \beta(F_O)$ is identified. Thus, asymptotically, the posterior distribution of β given the data in model $M = 1$ will be normal with mean $\beta\left(\widehat{F}_O\right)$ and variance equal to a consistent estimator of the asymptotic variance of $\beta\left(\widehat{F}_O\right)$.

11.2.3. Allowing for non-zero probability of non-confounding and non-causality. It is argued in Robins (1997b) that practicing epidemiologists will assign a prior probability of zero to the event of no confounding [i.e., the event that q is precisely zero] as in models $M = 0$ and $M = 1$ above. However, as discussed in Robins (1997b) and Robins and Wasserman (1998), the "faithfulness" analyses of Spirtes, Glymour, and Scheines (1993) and Pearl and Verma (1991) that allow one to go from association to causation without subject matter-specific knowledge rely on the assumption that there is a non-zero probability of non-confounding. Thus, to help further understand the results of Sprites et al. and Pearl and Verma, it is of interest to study the effect on our inferences of allowing a non-zero prior probability of non-confounding. Thus, we let model $M = 2$ be the model in which $q = r = 0$ *a priori* and $\pi[\eta \mid M = 2]$ is smooth. Note here that we have assumed that if $q = 0$, then we should choose $r = 0$ as well. with prior probability 1 since, if conditional on V, different levels of treatment are comparable with respect to the counterfactual Y_0, then it is reasonable to assume that they are comparable with respect to the magnitude of the treatment effect on a logistic scale. We let model $M = 3$ be the model in which $q = r = \gamma^* = 0$ and $\pi[\nu_- \mid M = 3]$ is smooth, where $\nu_- = \left(F_V, F_{A|V}, t\right)$ is the parameter ν less the component q. Again, we stress that we believe practicing epidemiologists will assign a prior probability of zero to the models $M = 2$ and $M = 3$. However, we will not do so here to help understand the results of Spirtes et al. and Pearl and Verma.

Model 2 is a non-parametric just-identified model for the law F_O. Thus the approximate Bayes factor $f(data \mid M = 2)/f(data \mid M = 0)$ comparing model 2 to model 0 is

$$\begin{aligned} &\pi_{F_O|M=2}\left(\widehat{F}_O\right)/\pi_{F_O|M=0}\left(\widehat{F}_O\right) \\ &= \frac{\pi[\widehat{\eta}(0) \mid M = 2]\,|\partial s(\widehat{\eta}(0), 0)/\partial\eta|^{-1}}{\int \pi[\widehat{\eta}(q), q, r \mid M = 0]\,|\partial s(\eta, q)/\partial\eta|^{-1}_{\eta=\widehat{\eta}(q)}\,dqdr} . \end{aligned} \tag{11.10}$$

In contrast to models 0–2, model 3 is no longer a non-parametric model for F_O. In fact, it imposes the sole restriction that Y is mean-independent of A given V. That is,

$$E[Y \mid A, V] = E[Y \mid V] . \tag{11.11}$$

Hence, under model 3, F_O lies in a space of dimension $d_A d_V + d_V - 1$ rather than a space of dimension $2d_A d_V - 1$. It follows, by Laplace's method. that

$$\begin{aligned} &f[data \mid M = 3] \propto \\ &f\left[data \mid \widehat{F}_{res}\right] n^{-\left(d_V d_A + d_V - 1\right)/2} \pi_{F_O|M=3}\left(\widehat{F}_{res}\right)\left[1 + O_p\left(n^{-1}\right)\right] \end{aligned} \tag{11.12}$$

where $\widehat{F}_{res}$ is the maximum likelihood estimator of F_O when (11.11) is imposed. Thus the Bayes factor $f[data \mid M = 3]/f[data \mid M = 0]$ is, to

$O_p\left(n^{-1}\right)$, given by the ratio of (11.12) to (11.9) evaluated at $m = 0$. If the model $M = 3$ is true, then this Bayes factor will tend to infinity at rate $O_p\left(n^{d_V(d_A-1)/2}\right)$. If model $M = 0$, $M = 1$, or $M = 2$ is true, then this Bayes factor will tend to zero exponentially quickly. [Note that following Sprites et al. and Pearl and Verma, we have assigned probability zero to any model for which the function $q(v,a)$ is not identically zero for all (v,a) and *(ii)* $q(v,a)$ is zero for some v and some a other than $a = 0$. Extensions of our results that do not impose this latter prior can easily be obtained.] It follows that if the prior probability $\pi(M = 3)$ that model 3 holds exceeds $O_p\left(n^{d_V(d_A-1)/2}\right)$ and (11.11) is true, we will asymptotically conclude that $q = \gamma^* = r = 0$ and thus conclude both no confounding and no causal effect of treatment. That is, we will have gone from association to causation without strong background subject matter knowledge.

We have only considered simple examples. In practice, we will be interested in examples where K is large and L_k and A_k can be multivariate with continuous and discrete components. It is a major open research question how to generalize the approach described above to this more complicated and realistic setting.

APPENDIX

A. Proof of Theorem 8.2. We shall need the following lemma.

LEMMA A.1. *Suppose* $Y = (Y_1, \ldots, Y_M)$ *given* A *has a continuous density w.r.t. Lebesgue measure on* R^M. *Then given a function* $q(y,a)$ *satisfying* $q(y,0) = 0$, *and a joint distribution for* (Y,A) *specified via densities* $f_{Y|A=a}(y) \equiv f_{Y|a}(y)$ *and* $f_A(a) \equiv f(a)$, *there exists a unique function* $\gamma(y,a)$ *satisfying (i)* $\gamma(y,0) = y$, *(ii) for each* a, $\gamma(y,a)$ *is a one-one function of* y, *(iii) if* $y_1 > y_2$, *then* $\gamma(y_1,a) > \gamma(y_2,a)$, *(iv)* $\gamma(y,a) \to \infty$ *as* $y \to \infty$ *and* $\gamma(y,a) \to -\infty$ *as* $y \to -\infty$, *and (v) with* $U \equiv \gamma(Y,A)$,

$$f[a \mid U = y] = t(a)\exp[q(y,a)] \Big/ \left\{ \int t(a)\exp[q(y,a)]\, d\mu(a) \right\} \tag{A.1}$$

for some $t(a)$ *satisfying* $\int t(a)\, d\mu(a) = 1$. *Specifically, with the function* $\gamma^{-1}(u,a)$ *defined by* $\gamma^{-1}(u,a) = y$ *if* $\gamma(y,a) = u$, $\gamma^{-1}(u,a)$ *is the unique solution to*

$$F_{Y|a}\left(\gamma^{-1}(u,a)\right) = \tau(u,a)/\tau(\infty,a) \tag{A.2}$$

where

$$\tau(u,a) = \int_{-\infty}^{u} f_{Y|0}(y)\exp[q(y,a)]\, dy\ . \tag{A.3}$$

Furthermore,

$$t(a) = f(a)\,\tau(\infty, a)^{-1} \Big/ \left\{ \int f(a)\,\tau(\infty, a)^{-1}\, d\mu(a) \right\} \tag{A.4}$$

and

$$f_U(y) = f_{Y|0}(y) \int_{-\infty}^{\infty} f(a) \exp[q(y, a)] \{\tau(\infty, a)\}^{-1}\, d\mu(a) \ . \tag{A.5}$$

Proof. Note (A.1) implies

$$f_{U|a}(y) / f_{U|a}(0) = \exp[q(y, a)]\, f_{U|0}(y) / f_{U|0}(0) \tag{A.6}$$

where, without loss of generality, we assume 0 is in the support of U. We now show that if $\gamma(y, a)$ satisfying *(i)-(v)* exists, then (A.2) must be true. Since, by *(i)*, $f_{U|0}(y) = f_{Y|0}(y)$, it follows from the change of variables formula, that (A.6) implies

$$\left|\partial \gamma^{-1}(y, a) / \partial y\right| f_{Y|a}\left(\gamma^{-1}(y, a)\right) = \exp\{q(y, a)\}\, f_{Y|0}(y) / k(a) \tag{A.7}$$

where $k(a) \equiv f_{U|a}(0) / f_{Y|0}(0)$. Upon integrating both sides of (A.7) over y in the set $[-\infty, u]$, we obtain

$$F_{Y|a}\left[\gamma^{-1}(u, a)\right] = \int_{-\infty}^{u} \exp[q(y, a)]\, f_{Y|0}(y)\, dy / k(a) \ . \tag{A.8}$$

Evaluating (A.8) as $u \to \infty$ and thus, by *(iii)*, as $\gamma^{-1}(u, a) \to \infty$, we obtain $k(a) = \int_{-\infty}^{\infty} \exp[q(y, a)]\, f_{Y|0}(y)\, dy$. Hence, $\gamma^{-1}(u, a)$ is given by (A.2), which has a unique solution since the RHS of (A.8) is a continuous multivariate distribution.

We next obtain (A.5) assuming *(i)-(v)*. Multiply both sides of (A.6) by $f_{U|a}(0)$ and integrate with respect to y to obtain

$$1 = \int_{-\infty}^{\infty} f_{U|a}(y)\, dy = f_{U|a}(0)\, \tau(\infty, a) / f_{U|0}(0) \ . \tag{A.9}$$

Hence,

$$f_{U|a}(0) = f_{U|0}(0) / \tau(\infty, a) \ . \tag{A.10}$$

Substituting the RHS of (A.10) for $f_{U|a}(0)$ in (A.6) and solving for $f_{U|a}(y)$, we obtain

$$f_{U|a}(y) = \exp[q(y, a)]\, f_{Y|0}(y) / \tau(\infty, a) \ . \tag{A.11}$$

Hence, $f_U(y) = \int f_{U|a}(y)\, f(a)\, d\mu(a)$ is given by (A.5).

We next obtain (A.4) under *(i)-(v)*. By (A.1) and Bayes' Theorem, we have

$$t(a) = \left\{f_{U|a}(y) f(a) / f_U(y)\right\} \exp\left[-q(y,a)\right] j(y) \tag{A.12}$$

with $j(y) = \int t(a) \exp\left[q(y,a)\right] d\mu(a)$. By $\int t(a)\, d\mu(a) = 1$, (A.12) implies $j(y)/f_U(y) = \left\{\int f_{U|a}(y) f(a) \exp\left[-q(y,a)\right] d\mu(a)\right\}^{-1}$. Hence, by (A.12),

$$t(a) = \exp\left[-q(y,a)\right] f_{U|a}(y) f(a) \Big/ \left\{\int \exp\left[-q(y,a)\right] f_{U|a}(y) f(a) d\mu(a)\right\}. \tag{A.13}$$

Upon substituting the RHS of (A.11) for $f_{U|a}(y)$ into the numerator and denominator of (A.13), we obtain (A.4).

We thus conclude that if *(i)-(v)* hold, then (A.2), (A.4), and (A.5) are true. Hence, the theorem is true if we can show that *(1.)* the density of U as given by (A.5) integrates to 1, *(2.)* $\gamma^{-1}(y,a)$ and $t(a)$ given by (A.2) and (A.4) satisfy *(i)-(v)*, and *(3.)* the change of variables formula

$$f_{Y|a}(y) f_A(a) = \left\{\partial\gamma(y,a)/\partial y\right\} f_U\left[\gamma(y,a)\right] f\left[a \mid U = \gamma(y,a)\right] \tag{A.14}$$

is satisfied with the RHS of (A.14) defined by (A.2), (A.4) and (A.5). It is straightforward to verify *(1.)-(3.)*. □

Proof of Theorem 8.2. We prove the theorem by a backward recursion beginning with K. Specifically, we apply Lemma A.1 conditional $\overline{L}_K = \overline{\ell}_K, \overline{A}_{K-1} = \overline{a}_{K-1}$. Then *(i)-(iv)* of Lemma A.1 are satisfied with $\gamma(y,a) \equiv \gamma_K\left(\underline{y}_K \overline{\ell}_K, \overline{a}_K\right), A \equiv A_K, Y \equiv \underline{Y}_{K+1}$. Further, it follows from model (8.3) and Theorem A.1 that conditional on $\overline{L}_K = \overline{\ell}_K, \overline{A}_{K-1} = \overline{a}_{K-1}$, *(v)* of Lemma A.1 holds with $U = U_K, q(y,a) \equiv q_K\left(\underline{y}_{K+1}, \overline{\ell}_K, \overline{a}_K\right)$ and $t(a) = t\left(a_K \mid \overline{\ell}_K, \overline{a}_{K-1}\right)$. Hence, we conclude from Lemma A.1 that model (8.3) is a non-parametric model for the law of the observed data given $\overline{L}_K = \overline{\ell}_K, \overline{A}_{k-1} = \overline{a}_{K-1}$ and, furthermore, that (8.4)–(8.6) are correct with $m = K$. To proceed we then reapply Lemma A.1 but now conditional on $\overline{L}_{K-1} = \overline{\ell}_{K-1}, \overline{A}_{K-2} = \overline{a}_{K-2}$. Specifically with $\gamma(y,a) \equiv \gamma_{K-1}\left(\underline{y}_K, \overline{\ell}_{K-1}, \overline{a}_{K-1}\right), A = A_{K-1}, Y = \left(Y_K, U_K'\right)'$, *(i)-(iv)* of Lemma A.1 hold. Further, again by Theorem 8.1, *(v)* of Lemma A.1 holds with $q(y,a) = q_{K-1}\left(y, \overline{\ell}_{K-1}, \overline{a}_{K-1}\right), U = U_{K-1}, t(a) = t\left(a_{K-1} \mid \overline{\ell}_{K-1}, \overline{A}_{K-2}\right)$. We thus obtain (8.4)–(8.6) for $m = K-1$ by applying Lemma A.1. We continue by backward recursion until $K = 0$. □

Proof of Theorem 8.3. To prove Theorem 8.3, we may use the following lemma.

LEMMA A.2. *Suppose $Y = (Y_1, \ldots, Y_M)$ given A has a continuous density with respect to Lebesgue measure on R^M. Given a function $\gamma(y,a)$ satisfying (i)-(iv) of Lemma A.1, and a joint density for (Y,A) specified by*

the densities $f_{Y|a}(y), f_A(a) = f(a)$, *there exists a unique function* $q(y,a)$ *satisfying* $q(y,0) = 0$ *such that A.1 holds with* $U = \gamma(Y,A)$. *Specifically,*

$$\exp[q(y,a)] = \{f_{U|a}(y)/f_{U|a}(0)\}\{f_{U|0}(0)/f_{U|0}(y)\}.$$

Proof. It follows immediately from the fact that A.1 implies A.6. □

Proof of Theorem 8.3. Theorem 8.3 follows by recursive application of Lemma A.2. The details are similar to that of the proof of Theorem 8.2 and are omitted. □

B. Proof of Theorem 8.14a. We shall prove the theorem for $\Phi(x) = x$. The proof for $\Phi(x) = e^x$ is similar. It is sufficient to show

$$E\left[g_k\left(\overline{A}_k, V_0^*\right) H_k\left(\psi^*, \eta^*\right) \mid V_0^*\right] = 0 . \tag{B.1}$$

By (8.53), (B.1) is equivalent to

$$\begin{aligned} 0 &= E\left[\left\{\sum_{j=0}^{k} m_{k+1}^*\left(\overline{L}_k, \overline{A}_k\right)\right\} g_k\left(\overline{A}_k, V_0^*\right) \Big/ \prod_{m=0}^{k} f\left(A_m \mid \overline{A}_{m-1}, L_m\right) \mid V_0^*\right] \\ &= \iint \prod_{m=0}^{k} d\mu(A_m)\, g_k\left(\overline{A}_k, V_0^*\right) \left[\sum_{j=0}^{k} \iint m_{k+1}^*\left(\overline{L}_j, \overline{A}_k\right)\right. \\ &\quad \left. \times \prod_{m=0}^{j} dF\left[L_m \mid \overline{L}_{m-1}\overline{A}_{m-1}, V_0^*\right]\right] . \end{aligned}$$

However, by (8.55), $\int m_{k+1}^*\left(\overline{L}_j, \overline{A}_k\right) dF\left[L_j \mid \overline{L}_{j-1}, \overline{A}_{j-1}, V_0^*\right] = 0$.

REFERENCES

BAKER, S.G., ROSENBERGER, W.F., AND DERSIMONIAN, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, **11**:643–657.

BALKE, A. & PEARL, J. (1997). Bounds on Treatment from Studies with Imperfect Compliance. *Journal of the American Statistical Association*, **92**:1171–1176.

BICKEL, P.J., KLAASSEN, C.A.J., RITOV, Y., AND WELLNER, J.A. (1993). **Efficient and Adaptive Inference in Semiparametric Models**. Baltimore, MD: Johns Hopkins University Press.

CHAMBERLAIN, G. (1987). Asymptotic Efficiency in Estimation with Conditional Moment Restrictions. *Journal of Econometrics*, **34**:305–324.

CORNFIELD, J., HAENSZEL, W., HAMMOND, E.C., LILIENFELD, A.M., SHIMKIN, M.B., AND WYNDER, E.L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, **22**:173–203.

DABROWSKA, D. (1988). Kaplan-Meier estimate on the plane. *Annals of Statistics*, **16**:1475–1489.

GILL, R.D. AND ROBINS, J.M. (1996). Sequential Models for Coarsening and missingness. **Proceedings of the First Seattle Symposium on Survival Analysis**, Springer-Verlag Lecture Notes in Statistics, pp. 295–305.

GILL, R.D., VAN DER LAAN, M.J., AND ROBINS, J.M. (1996). Coarsening at random: Characterizations, conjectures and counterexamples. **Proceedings of the First Seattle Symposium on Survival Analysis**, Springer-Verlag Lecture Notes in Statistics, pp. 255–294.

HECKMAN, J.J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Ann. Econ. Socl. Measurement.*, **5**:475–492.

HEITJAN, D.F., AND RUBIN, D.B. (1991). Ignorability and Coarse Data. *The Annals of Statistics*, **19**:2244–2253.

KLEIN, J.P. AND MOESCHBERGER, M.L. (1988). Bounds on net survival probabilities for dependent competing risks. *Biometrics*, **44**:528–538.

LIN, D.Y., PSATY, B.M., AND KRONMAL, R.A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, **54**:948–963.

LITTLE, R.J., AND RUBIN, D.B. (1987). **Statistical Analysis with Missing Data**. New York: John Wiley.

LITTLE, R.J.A. (1994), A Class of Pattern-Mixture Models for Normal Missing Data. *Biometrika*, **81**:471–483.

MANSKI, C.F. (1990). Nonparametric bounds on treatment effects. *American Economic Reviews, Papers, and Proceedings*, **80**:319–323.

MOESCHBERGER, M.L. AND KLEIN, J.P. (1995). Statistical models for dependent competing risks. *Lifetime Data Analysis*, **1**:195–204.

NEWEY, W.K. (1990). Semiparametric Efficiency Bounds. *Journal of Applied Econometrics*, **5**:99–135.

NEWEY, W.K., AND MCFADDEN, D. (1993). Estimation in Large Samples. **Handbook of Econometrics** (Vol. 4), D. McFadden and R. Engler, eds., Amsterdam: North Holland.

NORDHEIM, E.V. (1984). Inference from Nonrandomly Missing Categorical Data: An Example from a Genetic Study on Turner's Syndrome. *Journal of the American Statistical Association*, **7**:772–780.

PEARL, J., AND VERMA, T. (1991). A theory of inferred causation. In **Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference**. J.A. Allen, R. Fikes, and E. Sandewall, eds., pp. 441–452. San Mateo, CA: Morgan Kaufmann.

RITOV, Y., AND WELLNER, J.A. (1988). Censoring, Martingales, and the Cox Model. *Contemporary Mathematical Statistics Inf. Stochastic Procedures*, N.U. Prabhu, editor, American Mathematical Society, **80**:191–220.

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

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

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

ROBINS, J.M. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, **79**:321–34.

ROBINS, J.M., BLEVINS D, RITTER G, AND WULFSOHN M. (1992). *G*-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology*, **3**:319–336.

ROBINS, J.M., BLEVINS D, RITTER G, AND WULFSOHN M. (1993). Errata to *G*-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology*, **4**:189.

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. (1996). Locally efficient median regression with random censoring and surrogate markers. **Proceedings of the 1994 Conference on Lifetime Data Models in Reliability and Survival Analysis**, Boston, MA. In: **Lifetime Data: Models in Reliability and Survival Analysis**, N.P. Jewell et al., eds., Kluwer Academic Publishers, 263–274.

ROBINS, J.M. (1997a). Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine,* **16**:21–37.

ROBINS, J.M. (1997b). 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). Marginal structural models. In: *1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, pp. 1–10.

ROBINS, J.M. (1999b). Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference. **Statistical Models in Epidemiology, the Environment and Clinical Trials** , M. Elizabeth Halloran and Donald Berry, editors, NY: Springer-Verlag, pp. 95–134.

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

ROBINS, J.M. AND GILL, R. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine*, **16**:39–56.

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., AND ROTNITZKY, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In: **AIDS Epidemiology — Methodological Issues**. Jewell N., Dietz K. and Farewell V., eds., Boston, MA: Birkhäuser, pp. 297–331.

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 WASSERMAN, L. (1999). On the impossibility of inferring causation from association without background knowledge. **Computation, Causation, and Discovery**. C. Glymour and G. Cooper., eds., Cambridge, MA: The MIT Press (to appear).

ROSENBAUM, P.R. (1995). **Observational Studies**. New York: Springer-Verlag.

ROSENBAUM, P.R., AND RUBIN, D.B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B,* **11**:212–218.

ROTNITZKY, A., AND ROBINS, J.M. (1997). Analysis of semiparametric regression models with non-ignorable non-response. *Statistics in Medicine,* **16**:81–102.

ROTNITZKY, A., ROBINS, J.M. AND SCHARFSTEIN, D. (1998). Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association,*. **93**:1321–1339.

RUBIN, D.B. (1976). Inference and Missing Data. *Biometrika*, **63**:581–592.

SCHARFSTEIN, D., ROTNITZKY, AND A., ROBINS, J.M. (1999). Adjusting for non-ignorable drop-out with semiparametric non-response models (to appear, *Journal of the American Statistical Association*).

SCHLESSELMAN J.J. (1978). Assessing effects of confounding variables. American Journal of *Epidemiology*, **108**:3–8.

SLUD, E.V. AND RUBENSTEIN, L.V. (1983). Dependent competing risks and summary survival curves. *Biometrika,* **70**:643–649.

SPIRTES, P., GLYMOUR, C., AND SCHEINES, R. (1993). **Causation, Prediction, and Search**. New York: Springer Verlag.

VAN DER LAAN, M.J. AND ROBINS, J.M. (1998). Locally efficient estimation with current status data and time-dependent covariates. *Journal of the American Statistical*

Association, **93**:693–701.

VAN DER VAART, A. (1991). On differentiable functionals. *Annals of Statistics,* **19**:178–204.

ZHENG, M. AND KLEIN, J.P. (1994). A self-consistent estimator of marginal survival functions based on dependent competing risks and an assumed copula. *Communication in Statistics — Theory and Methods*, **23**:2299–2311.

ZHENG, M. AND KLEIN, J.P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, **82**:127–138.