Estimation in Missing Data Models

James Robins and Andrea Rotnitzky

Department of Biostatistics, Harvard School of Public Health

677 Huntington Avenue, Boston, MA 02115, U.S.A.

email: robins@hsph.harvard.edu

Introduction

The purpose of this paper is to discuss the advantages of using locally (semiparametric) efficient augmented inverse probability weighted (LEAIPW) estimators rather than fully parametric "model-based" estimators (which include parametric multiple-imputation estimators) in missing data problems (Robins, Rotnitzky,and Zhao, 1994). For example, when the only source of missing data is by design, and thus the non-response probabilities are known, LEAIPW estimators guarantee unbiasedness, while often vastly improving upon the poor efficiency of the estimator of Horvitz and Thompson. When missingness is unplanned rather than by design, LEAIPW estimators are still considerably more robust than parametric multiple-imputation or fully parametric likelihood-based estimators. If non-response is non-ignorable and we use a selection model parametrization, the consistency of a LEAIPW estimator only requires a correctly specified model for the non-response probabilities. In contrast, for consistency, a parametric model-based estimator additionally requires a correctly specified parametric model for the marginal distribution of the complete data. When non-response is ignorable, a LEAIPW estimator is doubly protected (i.e., doubly robust); that is, it is consistent if either a model for non-response or a parametric model for the complete data can be correctly specified. Therefore, an analyst can specify both a model for non-response and a parametric model for the complete data and his LEAIPW estimator will be consistent if he is lucky enough to have correctly specified either model. In contrast, the consistency of a parametric model-based estimator requires correct specification of a parametric model for the complete data.

We will next turn to the question of the ease with which doubly protected estimators can be computed in ignorable models. We will show that we can obtain closed form formulae for doubly robust semiparametric estimators whenever the missingness (non-response) process is a partial likelihood. In particular, this will be the case whenever the missing data pattern is monotone. Furthermore, we will show that there are many important non-monotone missing data problems in which the missingness process is a partial likelihood. In particular, we will show that if we wish to estimate the causal effect of a time-dependent treatment on an outcome and we can regard treatment as randomly assigned at each time given past treatment and covariate history, the resulting statistical problem can be viewed as a missing data problem with a partial-likelihood missingness mechanism, allowing for the construction of closed-form doubly robust estimators. That is, estimators that will be consistent for the causal effect of treatment if one correctly specifies either a model for the missingness (assignment) mechanism (i.e., the probability of treatment given the past treatment and covariate history) or a model for the remaining part of the likelihood. When the non-response process is not a partial likelihood, construction of doubly robust estimators generally requires solving integral equations which do not admit a closed form solution. Thus they are much more computationally demanding. Some of the remarks in this paragraph are unavoidably imprecise. In the following section we make them precise.

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 $Z(u), \overline{Z}(t) \equiv \{Z(u); 0 \le u \le t\}$. 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 $\overline{A} = \{A(u); 0 \le u < \infty\}$ with $A(u)$ taking values in a set $\mathcal{A}(u)$ and the existence of underlying counterfactual random variables

$$\{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\} \tag{1.1}$$

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

$$\overline{L}_{\overline{a}(u),0}(u) = \overline{L}_{\overline{a}(t),0}(u) = \overline{L}_{\overline{a}}(u) = \overline{L}_{\overline{a}^{\dagger}}(u) \tag{1.2}$$

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

$$\overline{L}(\tau) = \overline{L}_{\overline{A}(\tau),0}(\tau) \; . \tag{1.3}$$

Eq. (1.3) 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 $\overline{L}_{\overline{a}} = (\overline{Y}_{\overline{a}}, \overline{V}_{\overline{a}})$ 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) considers the sequential randomization assumption that for all t and $\overline{a} \in \overline{\mathcal{A}}$,

$$\underline{Y}_{\overline{a}}(t) \coprod A(t) \mid \overline{L}(t^{-}), \overline{A}(t^{-}) \tag{1.4}$$

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 (1.4) as the assumption of no unmeasured confounders given prognostic factors $L(t)$. Because of measurability issues, (1.4) 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., $\overline{L}(t^{-}) \equiv \lim_{u \uparrow t} \overline{L}(u)$, (1.4) is formally, for each t_k,

$$f\left[A(t_k) \mid \overline{L}(t_k^{-}), \overline{A}(t_k^{-}), \underline{Y}_{\overline{a}}(t_k)\right] = f\left[A(t_k) \mid \overline{L}(t_k^{-}), \overline{A}(t_k^{-})\right] \; . \tag{1.5}$$

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. (1.4) is formally that

$$\lambda_A\left[t \mid \overline{L}(t^{-}), \overline{A}(t^{-}), \underline{Y}_{\overline{a}}(t)\right] = \lambda_A\left[t \mid \overline{L}(t^{-}), \overline{A}(t^{-})\right] \tag{1.6a}$$

and

$$f[A(t) \mid \overline{L}(t^-), \overline{A}(t^-), A(t) \neq A(t^-), \underline{Y_{\overline{a}}}(t)] = \tag{1.6b}$$
$$f[A(t) \mid \overline{L}(t^-), \overline{A}(t^-), A(t) \neq A(t^-)] \ .$$

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. (1.6a) 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. (1.6b) 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 (1.4), 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

$$f[\overline{A}(\tau) \mid \{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\}] \ \textit{depends only on } O = (\overline{A}(\tau), \overline{L}(\tau)) \ . \tag{1.7}$$

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 $\{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\}$ one observes. Thus we can view causal inference as a missing data problem . We consider the following non-identifiable assumption concerning the statistical models for the full data $(\overline{A}, \{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\})$ considered in this section. Given $\overline{a}_1$ and $\overline{a}_2$, let u_{12} be the smallest time u with $a_1(u) \neq a_2(u)$.

Assumption A: For all $\overline{a}_1$ and $\overline{a}_2$ the conditional distribution of $(\overline{L}_{\overline{a}_1}, \overline{L}_{\overline{a}_2})$ 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 (1.4).

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 $L_{\overline{a}} = Y_{\overline{a}}$ process jumps only a $t = 1$, and that (1.4) 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 (1.7), nonetheless, it also implies $f[A(0)=0 \mid Y_1(1)] = 1 - q[Y_1(1)]$ which does violate (1.7) unless $Y_1(1) = Y_0(1)$ w.p.1, which is prohibited by Assumption A.

Lemma 1.1 has the following obvious partial converse if we strengthen (1.4).

Lemma: Suppose that (1.4) holds with $\{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\}$ replacing $\underline{Y}_{\overline{a}}(t)$. Then CAR holds.

However, if (1.4) is not so strengthened, then, even under Assumption A, the converse to Lemma 1.1 is not true. Specifically, Robins (1997, pg. 83) gives examples where one would expect (1.4) to be true even when (1.7) is false. However, if (1.7) 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 (1.4) is assumed, we can impose (1.7) without affecting our (non-parametric) inference. In the following remark, we show by counterexample that without Assumption A, CAR (i.e., 1.7) does not imply sequential randomization (i.e., 1.4), in which case the g-computation algorithm formula cannot be used to compute the marginal distribution of $\overline{Y}_{\overline{a}}$.

Remark 1.1: 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 $Y_{\overline{a}=(i,j)}$ [i.e., $\overline{Y}_{\overline{a}=(a(1^-),a(2^-))}$ 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

$$pr[A_1 = 1, A_2 = 0 \mid \{Y_{ij}; i,j = 1,2\}] = 1/8 + (1/8)\, Y_{10}(1) \tag{1.8a}$$

$$pr\left[A_1 = 0, A_2 = 1 \mid \{Y_{ij}; i,j = 1,2\}\right] = 1/8 + (1/8)\, Y_{01}(2) \tag{1.8b}$$

and

$$pr\left[A_1 = 1, A_2 = 1 \mid \{Y_{ij}; i,j = 1,2\}\right] = 3/4 - (1/8)\, Y_{10}(1) - (1/8)\, Y_{01}(2) \ . \tag{1.8c}$$

Now, by (1.2), $Y_{10}(1) = Y_{11}(1)$ 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 (1.8c) and check that the data are CAR. However, we now show that $pr\left[A_1 = 0 \mid Y_{10}(1)\right] \neq pr\left[A_1 = 0\right]$ in violation of (1.4). Specifically, $pr\left[A_1 = 0 \mid Y_{01}(2)\right]$ depends on $Y_{01}(2)$ by (1.8b). Furthermore, $pr\left[A_1 = 0 \mid Y_{10}(1)\right] =$
$pr\left[A_1 = 0, A_2 = 1 \mid Y_{10}(1)\right] = 1/8 + (1/8)\, E\left[Y_{01}(2) \mid Y_{10}(1)\right]$ 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 a: Assumptions concerning the joint distribution of $\left(\overline{L}_{\overline{a}_1}, \overline{L}_{\overline{a}_2}\right)$ given $\overline{L}_{\overline{a}_1}\left(u_{\overline{12}}\right)$ 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 1.1, 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 $\Omega(j) \equiv \int E\left[Y(2) \mid A_1 = j, A_2 = 1, Y(1)\right] dF\left[Y(1) \mid A_1 = j\right]$ 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 1.1 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 b: The example of Remark 1.1 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).

When (1.4) is true we can, as shown below, always obtain closed form doubly robust estimators. However if Assumption A is false but the data are CAR, these estimators will not include an efficient estimator and non-closed form efficient doubly robust estimators can be constructed. When CAR holds but (1.4) is false, then there exist only non-closed form doubly robust estimators. Here is an example in discrete time in which both Assumption A and CAR holds.

2. Discrete Time Marginal Structural Model

In this section, the temporally ordered observed data are $O = (L_1, A_1, L_2, A_2, ...L_K, A_K, L_{K+1})$ where A_k is a treatment given at time k and L_k are other variables measured just prior to treatment. Associated with each history $\overline{a} = (a_1, ..., a_K)$ there is a counterfactual random variable $L_{\overline{a}} = \overline{L}_{\overline{a},K+1}$ satisfying the consistency assumption $\overline{L}_{\overline{a},m} = \overline{L}_m$ if $\overline{A}_{m-1} = \overline{a}_{m-1}$. We impose the assumption of sequential ignorability (i.e., no unmeasured confounders) that *for all* $\overline{a}$ *and* m

$$L_{\overline{a}} \coprod A_m \mid \overline{L}_m, \overline{A}_{m-1} = \overline{a}_{m-1}. \tag{2.1}$$

so that CAR holds. Further we assume that for all a_m in the support of A_m

$$\begin{aligned} &\textit{If } f\left(\overline{A}_{m-1}, \overline{L}_m\right) > 0, \\ &\textit{then } f\left[a_m \mid \overline{A}_{m-1}, \overline{L}_m\right] > 0. \end{aligned} \tag{2.2}$$

Robins (1999) discusses the use of MSMs to estimate the effect of $\overline{a}$ on $L_{\overline{a}}$. With little loss of generality, he shows that the statistical problem which arises from estimation of MSMs under sequential ignorability reduces to the estimation of the finite dimensional parameter μ in a semiparametric model with likelihood $f(O;\mu,\theta,\rho) = \prod_{m=1}^{K+1} f\left[L_m \mid \overline{L}_{m-1}, \overline{A}_{m-1}; \mu, \theta\right] \times$ $\prod_{m=1}^{K} f\left[A_m \mid \overline{A}_{m-1}, \overline{L}_m; \rho\right]$, where
$f\left[L_m \mid \overline{L}_{m-1}, \overline{A}_{m-1}; \mu, \theta\right]$ and
$f\left[A_m \mid \overline{A}_{m-1}, \overline{L}_m; \rho\right]$ are densities with respect to dominating measures ν_l and ν_a respectively, (μ, θ) and ρ are variation-independent, and θ and ρ are infinite-dimensional nuisance parameters, that is characterized by the restriction for all functions $s = s(\cdot,\cdot,\cdot)$ in a given set $\mathcal{S}$

$$E\left[s\left(\overline{L}_{K+1}, \overline{A}_K; \mu\right)/\overline{\pi}_K\right] = 0 \tag{2.3}$$

where $\overline{\pi}_m = \prod_{j=1}^{m} f\left[A_j \mid \overline{L}_j, \overline{A}_{j-1}\right]$. Note that the missingness process $\prod_{m=1}^{K} f\left[A_m \mid \overline{A}_{m-1}, \overline{L}_m; \rho\right]$ is a partial likelihood. Specifically, if we specify a marginal structural model characterized by the restriction that $E\{s(L_{\overline{a}}, \overline{a}; \mu)\} = 0$ for $s \in \mathcal{S}$, then (2.3) will hold under (2.1) and (2.2). As an example, the marginal structural mean model $E\left[L_{\overline{a},K+1}\right] = g(\overline{a}, \mu)$ with $g(\cdot,\cdot)$ a known function is equivalent to the model $E\{s(L_{\overline{a}}, \overline{a}; \mu)\} = 0$ with $\mathcal{S} = \{s(L_{\overline{a}}, \overline{a}; \mu) = s^*(\overline{a})\left[L_{\overline{a},K+1} - g(\overline{a}, \mu)\right]; s^*(\cdot)$ *arbitrary*$\}$. If the assumption that $E\{s(L_{\overline{a}}, \overline{a}; \mu)\} = 0$ for $s \in \mathcal{S}$ does not restrict the distribution of the $L_{\overline{a}}$,we say our MSM is saturated.

In order to reduce dimensionality, we consider two parametric submodels, $f\left(a_m \mid \overline{l}_m, \overline{a}_{m-1}; \alpha\right)$ and $f\left(l_m \mid \overline{l}_{m-1}, \overline{a}_{m-1}; \mu, \psi\right)$ where ψ and α are finite-dimensional parameters. Our goal will be to construct a doubly-robust RAL estimators of μ in the semiparametric *union* model that assumes (2.1), (2.2), the model $f(O;\mu,\theta,\rho)$ (so (2.3) holds), and at least one of the two finite-dimensional submodels is correct as well. Further, when both parametric models are correct and our MSM is saturated, the estimator should attain the semiparametric variance bound for the union model. For simplicity we take μ to be one dimensional. In that case S will be a one dimensional vector space when the MSM is saturated. An estimator that satisfies our goal can be constructed as follows.

1. Compute the restricted MLE $\widehat{\psi}=\widehat{\psi}(\mu)$ of ψ with μ held fixed and the MLE $\widehat{\alpha}$ of α from the observed data.

2. Select a particular s from $\mathcal{S}$. (The choice can only affect efficiency.)

3. $\widehat{\mu}_{doublerobust}$ solves the augmented inverse probability of treatment weighted estimating equation $0=\sum_i U_i(\mu)$, where $U(\mu)=s\left(\overline{L}_{K+1},\overline{A}_K,\mu\right)/\overline{\pi}_K(\widehat{\alpha})-\sum_{m=0}^{K}\{c\left(m,\overline{L}_m,\overline{A}_m,\mu\right)-E_{\widehat{\alpha}}\left[c\left(m,\overline{L}_m,\overline{A}_m,\mu\right)|\overline{L}_m,\overline{A}_{m-1}\right]\}$ with $c\left(m,\overline{L}_m,\overline{A}_m,\mu\right)=E_{\mu,\widehat{\psi}(\mu)}\left[s\left(\overline{L}_{K+1},\overline{A}_K,\mu\right)/\overline{\pi}_K(\widehat{\alpha})|\overline{L}_m,\overline{A}_m\right]$.

   We have acted as if it were no problem to specify and maximize the fully parametric MSM likelihood. This is not the case as discussed in Sec. 8 of Robins et al. (1999) where variation independent parameterization of the observed data distribution are given for MSMs and are shown to be quite complex. An alternative parametrization would be to specify a parametric model depending on (μ,ψ) for the joint distribution of the counterfactuals $L_{\overline{a}}$, but this may lead to intractable high or even infinite dimensional integrals. The message is that the analysis of fully parametric MSMs is anything but straightforward. Thus the method of estimator construction given above may be more theoretical than practical because of the difficulty with fitting the required parametric models. In practice, we would replace $c\left(m,\overline{L}_m,\overline{A}_m,\mu\right)=E_{\mu,\widehat{\psi}(\mu)}\left[s\left(\overline{L}_{K+1},\overline{A}_K,\mu\right)/\overline{\pi}_K(\widehat{\alpha})|\overline{L}_m,\overline{A}_m\right]$ computed under the model by the predicted value $\widehat{c}\left(m,\overline{L}_m,\overline{A}_m,\mu\right)$ from a flexible regression of the variable $s\left(\overline{L}_{K+1},\overline{A}_K,\mu\right)/\overline{\pi}_K(\widehat{\alpha})$ on functions of $\overline{L}_m,\overline{A}_m$.

References

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

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 (9-12):923-945.

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.

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.

Robins J.M. (1999). Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference. In: **Statistical Models in Epidemiology: The Environment and Clinical Trials.** Halloran, M.E. and Berry, D., eds. NY: Springer-Verlag, pp. 95-134.

Robins JM, Scharfstein D, Rotnitzky A. (1999). Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. In: **Statistical Models in Epidemiology: The Environment and Clinical Trials.** Halloran, M.E. and Berry, D., eds. NY: Springer-Verlag, pp. 1-94.

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

Robins, J.M. (1997). Causal inference from complex longitudinal data. In: **Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics (120)**, M. Berkane, Editor. NY: Springer Verlag, 69-117.