

CORRECTION FOR NON-COMPLIANCE IN EQUIVALENCE TRIALS

JAMES M. ROBINS*

Epidemiology and Biostatistics Departments, Harvard School of Public Health, Boston, MA 02115, U.S.A.

SUMMARY

In randomized trials comparing a new therapy to standard therapy, the sharp null hypothesis of equivalent therapeutic efficacy does not imply the intent-to-treat null hypothesis of equal outcome distributions in the two-treatment arm if non-compliance is present. As a consequence, the development of analytic methods that adjust for non-compliance is of particular importance in equivalence trials comparing a new therapy to standard therapy. This paper provides, in the context of equivalence trial, a unified overview of various analytic approaches to correct for non-compliance in randomized trials. The overview focuses on comparing and contrasting the plausibility, robustness, and strength of assumptions required by each method and their programming and computational burdens. In addition, several new structural (causal) models are introduced: the coarse structural nested models, the non-nested marginal structural models and the continuous-time structural nested models, and their properties are compared with those of previously proposed structural nested models. The fundamental assumption that allows us to correct for non-compliance is that the decision whether or not to continue to comply with assigned therapy at time t is random (that is, ignorable or explainable) conditional on the history up to t of measured pre- and time-dependent post-randomization prognostic factors. In the final sections of the paper, we consider how the consequences of violations of our assumption of conditionally ignorable non-compliance can be explored through a sensitivity analysis. Finally, the analytic methods described in this paper can also be used to estimate the causal effect of a time-varying treatment from observational data. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

The development of analytic methods that adjust for non-compliance is of particular importance in randomized trials comparing a new therapy to standard therapy, since the sharp null hypothesis of equivalence does not imply the intent-to-treat null in the presence of non-compliance. To fix ideas, consider a randomized equivalence trial conducted by the manufacturer of new drug therapy in which the new therapy is to be compared to the standard proven therapy for a disease. Suppose all subjects are initially compliant. However, 50 per cent of subjects randomized to standard therapy and 20 per cent of subjects randomized to the new therapy later become non-compliant and stop all therapy due to mild, easily palliated side-effects. Suppose

* Correspondence to: James M. Robins, Professor of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.

Contract grant sponsor: National Institutes of Health

Contract grant numbers: 2P30 ES00002, R01-A132475, R01-ES03405, K04-ES00180, GM-48704, GM-29745

that, in fact, the new and the standard therapy are equally efficacious and both are much superior to being off all therapy. Then an intent-to-treat analysis that compares outcomes among subjects randomized to the new therapy with those randomized to standard therapy will indicate a beneficial effect of the new therapy that will be wholly attributable to the high non-compliance rate in the standard therapy arm. However, in the community, this high non-compliance rate might be avoided by making the palliative therapy necessary to alleviate side-effects easily available. Indeed, in this setting, even if the new therapy were less efficacious than standard therapy, the intent-to-treat analysis would suggest a beneficial effect of the new therapy, provided this therapy were much superior to no therapy. Thus, it is important to develop alternatives to the standard intent-to-treat analysis that can correct for non-compliance in equivalence trials.

The goal of this paper is to provide, in the context of a equivalence trial, a unified overview of various analytic approaches that I and co-workers have proposed to correct for non-compliance in randomized trials:¹⁻¹⁵ G-computation algorithm estimators; inverse probability of censoring weighted estimators; iterated conditional expectation estimators; G-estimation and likelihood-based estimation of rank-preserving structural models; structural nested distribution models; and structural nested mean models. Others have also contributed to the development of the above methods: Heyting *et al.*¹⁸ independently proposed inverse probability of censoring weighted estimators and Goetghebeur and co-authors have studied structural nested mean models in reference 39 and in several unpublished manuscripts. The overview focuses on comparing and contrasting (i) the plausibility, robustness, and strength of assumptions required by each method, and (ii) the programming and computation burdens. Proofs and theoretical details are, in the main, neglected; the reader is referred to previous publications for these. In addition to this unified overview, several new structural (causal) models are introduced: the coarse structural nested models, the non-nested marginal structural models and the continuous-time nested structural models. Their properties are compared with those of previously proposed structural nested models. The marginal structural models will be particularly useful when the outcome variable is dichotomous. The analytic approaches described in this paper rely on the fundamental assumption of explainable non-random non-compliance by measured time-dependent prognostic factors. I and co-workers have previously referred to this assumption as the assumption of no unmeasured confounders^{6,20} or the assumption of sequential or maintained randomization.^{2,10} It is a sequential version of Rosenbaum and Rubin's²² strong ignorability assumption. In Section 9 I present new methods for assessing the sensitivity of one's inferences to this assumption.

Owing to space limitations, there are several important topics, discussed in my previous work, that I only briefly consider here. For example, I consider estimating the outcomes that would have been observed had subjects followed their assigned treatment protocol. However, I only briefly consider the important question of estimating the outcomes that would have been observed under protocols other than those assigned. The interested reader is referred to references 1, 3 and 8, App. 1, 9, 12. The use of the treatment arm (randomization) indicator as an instrumental variable is an important analytic strategy for correcting for non-compliance in trials comparing either several levels of a single active treatment or a single active treatment to a placebo.^{3,4,10,16,17,19,21,23,35,39,40} This would also be an important analytic strategy in a equivalence trial in which there is treatment cross-over (that is, non-compliant subjects in one arm adopt the treatment assigned to the other arm).^{8,9,12} In this paper, we assume no treatment cross-over and do not use the randomization indicator as an instrument. Rather, we estimate the treatment-arm specific outcome distribution that would be observed under complete

compliance separately within treatment arms, and, then, as the final step in the analysis, compare these two distributions with one another. We view the within treatment arm analyses as 'observational study' analyses since they ignore the fact that the assignment to treatment arm was at random. Indeed, the same analytic methods can be used to estimate causal effects of a time-varying treatment from observational data. Again, due to space limitations, the paper contains no real data set or other worked example. However, references 1, 7–11, 13, 14 and 20 use these methods to reanalyse data from both randomized trials with non-random non-compliance and longitudinal observational studies.

2. ESTIMATION OF TREATMENT EFFECTS IN A EQUIVALENCE TRIAL

2.1. A simplified trial

In the equivalence trial discussed in Section 1, to fix notation, for subject $i = 1, \dots, n$, let: Z_i denote the dichotomous randomization indicator, that is, treatment arm; D_i denote the observed treatment; $Y_i(z, d)$ denote the outcome of interest that would be observed if (possibly contrary to fact) subject i were randomized to arm z and treatment d were given; $z = 0, 1, d = 0, 1, 2, 3$; and $Y_i = Y_i(Z_i, D_i)$ denote the observed outcome. Here the four possible realizations of D_i correspond to the four treatments: always remain on standard therapy ($D_i = 0$); always remain on the new therapy ($D_i = 1$); begin standard therapy, then stop all therapy ($D_i = 2$); begin the new therapy, then stop all therapy ($D_i = 3$). Our notation incorporates Rubin's SUTVA assumption²⁴ that the outcomes for a subject are not affected by the treatments or outcome of others. Throughout we shall assume the exclusion restriction that, given treatment, randomization arm z does not influence outcome, so $Y_i(d) \equiv Y_i(z, d)$ and $Y_i \equiv Y_i(D_i)$. By Z randomized, we also have that Z_i is jointly independent of the $Y_i(d)$, that is

$$Z_i \perp\!\!\!\perp \{Y_i(d); d = 1, 2, 3, 4\}$$

where $A \perp\!\!\!\perp B | C$ means A is independent of B given C . We shall assume that $(Z_i, D_i, \{Y_i(d); d = 0, 1, 2, 3\})$, $i = 1, \dots, n$ are i.i.d. realizations of a random vector. Given this notation, if side-effects are easily palliated, the sharp null equivalence hypothesis of interest is

$$Y_i(1) = Y_i(0) \quad \text{for all } i \quad (1)$$

which does not imply the ITT null that Y_i is independent of Z_i . The ITT null is implied by the sharp null hypothesis that both (1) and

$$Y_i(d) = Y_i \text{ in treatment arm } Z = d, d = 0, 1. \quad (2)$$

Equation (2) implies that being off therapy was as efficacious as being on the standard proven therapy in arm $Z = 0$ and is efficacious as being on the new therapy in arm $Z = 1$ (e.g., $Y_i(0) = Y_i(2)$ in arm $Z = 0$). Thus hypothesis (2) will not be of interest if standard therapy were already known to be superior to no therapy. If standard therapy were not already known to be superior, then the joint null hypothesis (1)–(2) and thus tests of the ITT null would be of interest. In this paper, we shall primarily be concerned with trials in which standard therapy is known to be superior to no therapy, although, for the sake of completeness, we will discuss the operating characteristics of our correction procedures under the joint null (1)–(2).

Throughout we shall assume the outcome Y_i is a univariate outcome measured at the end of the trial, say, 4 months post-randomization. This is solely to allow us to focus on the conceptual

issues that must be faced in correcting for non-compliance without the added technical complications of dealing with failure time and/or repeated measures outcomes. Co-workers and I have discussed methods for correcting for non-compliance in randomized studies with failure time and repeated measures outcomes in references 1–14. Until Section 10, we shall assume that there is no loss to follow-up, so the outcome Y_i is observed for each study subject. To keep technical complications to a minimum, until Section 6.2 we shall take as our parameter of interest the average treatment effect (ATE) parameter $E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$. The ATE parameter is not identifiable from equivalence trial data. For Y_i dichotomous, sharp bounds for the ATE parameter are $-1 + \text{pr}(Y_i = 1, D_i = 1 | Z_i = 1) + \text{pr}(Y_i = 0, D_i = 0 | Z_i = 0) \leq \text{ATE} \leq 1 - \text{pr}(Y_i = 0, D_i = 1 | Z_i = 1) - \text{pr}(Y_i = 1, D_i = 0 | Z_i = 0)$.^{23, 25} These bounds are the same whether or not Y_i is observed for subjects for whom D_i is 2 or 3 (that is, for subjects who stop therapy). That is, we can, and until Section 4, do regard a subject who stops therapy as ‘censored’ at that time, because that subject’s outcome data will not affect our inference and thus need not be recorded for data analysis.

If one wishes to identify the ATE parameter, further strong non-identifiable assumptions must be added. Co-workers and I^{1, 2, 5} and Heyting *et al.*¹⁸ have studied identifying assumptions for ATE in this setting. If the decision to quit therapy is essentially a second randomization, that is, $Y_i(d) \perp\!\!\!\perp D_i | Z_i$ for $d = 0, 1$, we say that the non-compliance is random where $A \perp\!\!\!\perp B | C$ means A is independent of B given C . In that case, $E[Y_i(d)]$ is identifiable and equal to $E[Y_i | Z_i = d, D_i = d]$ for $d = 0, 1$.

If, as is usually the case, one does not believe compliance is random given Z_i , one can try to collect data on additional pre- or post-randomization covariates L such that compliance is random conditional on the covariates, that is, $Y_i(d) \perp\!\!\!\perp D_i | Z_i, L_i, d = 0, 1$. We refer to this assumption as the assumption of explainable non-random non-compliance, because, within treatment arm Z , the non-randomness of the non-compliance is completely explained by the covariates L . As an ‘observational study-like’ assumption, the assumption of explainable non-random non-compliance is not testable from data and would never be expected to be precisely true. However, careful consideration in choosing and measuring the covariates L may result in the assumption of explainable non-random non-compliance being approximately true. Under this assumption, $E[Y_i(d)]$ is identifiable and, for a discrete post-randomization L , equals $\sum_{\ell} E[Y_i | Z_i = d, D_i = d, L_i = \ell] \text{pr}[L_i = \ell | Z_i = d]$ for $d = 0, 1$. Robins^{1, 2} called this formula the G-computation algorithm formula given covariates L .

Remark: If L_i is a pre-randomization covariate, then the ATE parameter $E[Y_i(1)] - E[Y_i(0)]$ is the weighted average $\sum_{\ell} \text{RD}_{\ell} \text{pr}[L_i = \ell]$ of the ℓ -stratum specific risk differences $\text{RD}_{\ell} = E[Y_i | D_i = 1, L_i = \ell] - E[Y_i | D_i = 0, L_i = \ell]$. (Here we have used the fact that, in this trial, $D_i = d$ implies $Z_i = d$ for $d = 0, 1$.) In contrast, if L_i is a post-randomization covariate affected by initial treatment so that $Z_i \not\perp\!\!\!\perp L_i$, then the ATE parameter cannot be written as a weighted average of ℓ -stratum specific risk differences!

In general, the post-randomization L_i may not be discrete. The general form of the G-computation algorithm formula is then given by $\int E[Y_i | Z_i = d, D_i = d, L_i = \ell] dF(\ell | Z_i = d)$. This formula can also be written as the inverse probability of censoring weighted (IPCW) estimand $E[Y_i I(Z_i = d, D_i = d) / \pi_d(1) \pi_{di}(2)]$ where $\pi_d(1) = \text{pr}[Z_i = d]$ and $\pi_{di}(2) \equiv \text{pr}[D_i = d | Z_i = d, L_i]$.^{5, 13, 18} The G-computation algorithm formula can also be written as the iterated conditional expectations (ICE) estimand $E[E\{Y_i | Z_i = d, D_i = d, L_i\} | Z_i = d]$ (reference 13, p. 120).

The IPCW estimand is easily generalized to allow investigation of the sensitivity of $E[Y_i(d)]$ to the fundamental assumption $Y_i(d) \perp\!\!\!\perp D_i | Z_i, L_i$ of explainable non-random non-compliance. Let $\pi_{di}(2, y) = \text{pr}[D_i = d | Z_i = d, L_i, Y_i(d) = y]$ be the probability of treatment $D = d$ given $Z_i = d, L_i$, and $Y_i(d) = y$. Note $\pi_{di}(2, y)$ is not identifiable since we do not observe $Y_i(d)$ for subjects for whom $D_i \neq d$. In a sensitivity analysis, we could select plausible functions $\pi_{di}(2, y)$ based on our prior beliefs. For a given $\pi_{di}(2, y)$, $E[Y_i(d)]$ is given by the IPCW estimand with $\pi_{di}(2, Y_i)$ substituted for $\pi_{di}(2)$. More generally, we specify a parametric model for the unknown $\pi_{di}(2, Y_i)$ and simultaneously estimate the parameters of that model and $E[Y_i(d)]$ (see Section 10 and references 41, 43, 13, p. 118, 27 and 28). However, the parameters of the non-ignorable missingness model and $E[Y_i(d)]$ may not be simultaneously identifiable. In that case, one can perform a sensitivity analysis by using the methods described in Section 9.2.

2.2. A more realistic trial

Realistically, different subjects will abandon therapy at different times. Suppose we record whether subjects are on their assigned therapy at fixed times $0, 1, \dots, K$ and Y_i is measured at time $K + 1$, and time zero is time of randomization.

Remark: The time interval between k and $k + 1$ does not have to correspond to successive clinic visits. Indeed it will often be advantageous to choose the interval to be one day. Since the timing of events is usually recorded only to the nearest day, this choice will result in no additional loss of information from grouping.

Until Section 10, we assume that Y_i is always observed. Let $R_{ki} = 0$ if a subject is on his assigned therapy in the time interval $(k, k + 1]$ and $R_{ki} = 1$ if the subject is off all therapy. Subjects are no longer assumed to initially be on their assigned therapy, but, until Section 8, we continue to assume subjects either are on their assigned therapy or are off all therapy. Note that for a subject in treatment arm $Z_i = d$, Y_i is equal to $Y_i(d)$ if the subject remains on therapy $\bar{R}_{Ki} = \mathbf{0}$ where we introduce the notation that for any variable X_k , $\bar{X}_k = (X_0, X_1, \dots, X_k)'$. In addition to R_{ki} , we record data on sufficient covariates $L_{0i}, L_{1i}, \dots, L_{Ki}$ at time $0, \dots, K$ such that (hopefully) to a good approximation, the decision to remain on therapy at time k is independent of the counterfactual variable $Y_i(d)$ conditional on past covariate history L_{ki} and treatment arm Z_i among subjects who have remained on the assigned therapy prior to time k (i.e., $\bar{R}_{(k-1)i} = \mathbf{0}$). That is,

$$R_k \perp\!\!\!\perp Y(d) | \bar{R}_{k-1} = \mathbf{0}, \bar{L}_k, Z = d, d = 0, 1 \quad (3)$$

where we have suppressed the i subscript (for example, $Y(d)$ is $Y_i(d)$) and $\mathbf{0}$ is a vector with each component 0. We refer to (3) as the assumption of (sequential) explainable non-random non-compliance. It is not testable from the data. Assumption (3) has been called the assumption of randomization with respect to Y for treatment regime $\bar{r}_K \equiv \mathbf{0}$ in reference 2. It is related to the assumption of no unmeasured confounders in reference 20 and to a sequential version of Rosenbaum and Rubin's strong ignorability assumption.²² We discuss how to assess the sensitivity of our inferences to this assumption in Section 9.

Remark: When the interval between times k and $k + 1$ is one day, we do not assume that the covariates L_k are measured daily. Rather, we adopt the convention that we assign to day k the last measured value of each component of L_k . The reason for this convention is that this is the value of

the covariate that is available to the physician and most likely to the patient on day k . If so, we might hope that is the value of the covariate on which treatment decisions are made on day k , so that the assumption (3) might be approximately true.

Under (3), $E[Y(d)]$ is identified and is given by the G-computation algorithm formula $\int \dots \int E[Y|\bar{\ell}_K, \bar{R}_K = \mathbf{0}, Z = d] \prod_{m=0}^K dF(\ell_m|\bar{\ell}_{m-1}, \bar{R}_{m-1} = \mathbf{0}, Z = d)$.^{3,36} That is $E[Y(d)]$ is a weighted average of the conditional mean of Y given $\bar{\ell}_K$ among subjects who remain on treatment in arm $Z = d$ (that is, $E[Y|\bar{\ell}_K, \bar{R}_K = \mathbf{0}, Z = d]$) with $\bar{\ell}_K$ specific weights $w(\bar{\ell}_K, d) \equiv \prod_{m=0}^K f(\ell_m|\bar{\ell}_{m-1}, \bar{R}_{m-1} = \mathbf{0}, Z = d)$.

The G-computation algorithm formula can be written as the IPCW estimand $E[Y\delta_d/\bar{\pi}_d(K)|Z = d]$ where $\delta_d = I(\bar{R}_K = \mathbf{0}, Z = d)$, $\bar{\pi}_d(k) = \prod_{m=0}^k \pi_d(m)$, $\pi_d(m) = \text{pr}[R_m = 0 | \bar{R}_{m-1} = \mathbf{0}, \bar{L}_m, Z = d]$. The IPCW estimand is a weighted sum of the values of Y among subjects in arm $Z = d$ who remained on treatment. Under a slight strengthening of the assumption (3) of sequential explainable non-random non-compliance, the weights $1/\bar{\pi}_d(K)$ are equal to the reciprocal of the probability of having remained on treatment (references 5, 26, 13, p. 109 and 18).

The G-computation algorithm formula can also be written as the iterated conditional expectations estimand (ICE), $\eta_d = E[\eta_{d0}|Z = d]$, where for $k = K - 1, \dots, 0$, η_{dk} is defined recursively by $\eta_{dk} = E[\eta_{d(k+1)}|\bar{L}_k, \bar{R}_k = \mathbf{0}, Z = d]$ with $\eta_{dK} \equiv E[Y|\bar{L}_K, \bar{R}_K = \mathbf{0}, Z = d]$. It can be shown that η_{dk} is the mean of $Y(d)$ given $\bar{L}_k, \bar{R}_k = \mathbf{0}$ in arm $Z = d$ under the explainable non-random non-compliance assumption (3) (reference 13, p. 120).

2.3. Conditions for Random Non-compliance

When (3) holds, if the covariate history \bar{L}_k does not predict R_k among subjects on their assigned therapy in arm $Z = d$, that is

$$R_k \perp\!\!\!\perp \bar{L}_k | \bar{R}_{k-1} = \mathbf{0}, Z = d, d = 0, 1 \tag{4}$$

then compliance will be random in the sense that R_k will be independent of $Y(d)$ among subjects remaining on therapy in arm $Z = d$, that is

$$R_k \perp\!\!\!\perp Y(d) | \bar{R}_{k-1} = \mathbf{0}, Z = d, d = 0, 1. \tag{5}$$

It would often be unlikely, based on substantive considerations, to believe that (5) were true unless one also believed that both (3) and (4) were true (reference 13, p. 109). Equation (4) is an assumption about the distribution of the observable random variables and thus is subject to empirical testing. In contrast, (3) is non-identifiable and thus not subject to empirical tests.

ESTIMATION

In this section, we consider estimation of $E[Y(d)]$ under the explainable non-random non-compliance assumption (3). In a realistic study the number of time intervals K may be as large as 1000 (if data are recorded daily), L_k may be highly multivariate with both continuous and discrete components, and Y may be continuous or discrete. If the sample size n is of the order of 500 to 1000, we shall require modelling assumptions (*a priori* restrictions on the joint distribution of the data) to estimate $E[Y(d)]$, $d = 0, 1$. We shall consider four different approaches to the estimation of $E[Y(d)]$ corresponding to four different ways of modelling the distribution of the observed data. The first three approaches are based on modelling the unknown components of the G-computation algorithm estimand, the IPCW estimand, and the ICE estimand. The fourth

method is a likelihood-based method that is based on modelling the marginal distribution of $Y(d)$. We first describe the four approaches to estimation, and then compare their relative strengths and weaknesses.

We consider estimation of $E[Y(d)]$ based on data from subjects in treatment arm $Z = d$. For notational convenience, we shall, henceforth, suppress the dependence on the treatment arm d . We shall also suppress dependence of conditioning events on the event that treatment arm $Z = d$.

3.1. Approach 1: G-computation Algorithm Estimator

We specify fully parametric models $f(L_m | \bar{L}_{m-1}, \bar{R}_{m-1} = \mathbf{0}; \theta_1)$ and $f(Y | \bar{L}_K, \bar{R}_K = \mathbf{0}; \theta_2)$ for the conditional laws of L_m and Y among subjects remaining on treatment in treatment arm $Z = d$ and then estimate $\theta = (\theta_1, \theta_2)$ by maximum likelihood using data from subjects in arm $Z = d$, that is, by $\hat{\theta} \sim (\hat{\theta}_1, \hat{\theta}_2)$ that maximizes $\prod_{i=1}^{n_d} \mathcal{L}_{gi}(\theta_1, \theta_2)$ where n_d is the number of subjects in arm $Z = d$,

$$\mathcal{L}_g(\theta_1, \theta_2) = \prod_{m=0}^{\min(C, K)} f[L_m | \bar{L}_{m-1}, \bar{R}_{m-1} = \mathbf{0}; \theta_1] f[Y | \bar{L}_K, \bar{R}_K = \mathbf{0}; \theta_2]^{I(C=K+1)}$$

where the ‘censoring time’ C is the first time k that a subject is off his/her assigned therapy ($R_k = 1$). By convention, $L_{K+1} \equiv 1$ and $R_{K+1} \equiv 1$ so $C = K + 1$ for a subject who remains on therapy throughout. The MLE $\hat{\beta}_g$ of $E[Y(d)]$ is then given by the G-computation algorithm formula evaluated at the maximum likelihood estimate of θ . In practice, when K is large, the formula must be evaluated by Monte Carlo integration.³ One important exception to the need to evaluate the formula by Monte Carlo integration is when the estimate of $E[Y(d)]$ is given by the generalized sweep estimator described in reference 13, p. 115. Worked examples using this method can be found in references 1 and 34.

3.2. Approach 2: IPCW Estimator

We specify a model for continuing compliance at time m give \bar{L}_m

$$\text{logit } \pi(m, \alpha) = \alpha_{1m} + \alpha'_2 h_m(\bar{L}_m) \quad (6)$$

for $\pi(m) \equiv \text{pr}[R_m = 0 | \bar{R}_{m-1} = \mathbf{0}, \bar{L}_m]$ among subjects in arm $Z = d$ where $h_m(\bar{L}_m)$ is a known vector valued function of \bar{L}_m ; $\alpha' = (\alpha'_1, \alpha'_2)$, $\alpha'_1 = (\alpha_{10}, \dots, \alpha_{1K})$, $m = 0, \dots, K$, and $\bar{R}_{-1} \equiv \mathbf{0}$ by convention. Note if (4) holds so compliance is completely at random (that is, (5) holds), then the compliance model (6) is correctly specified with $\alpha_2 = 0$.

The maximum likelihood estimator $\hat{\alpha}$ maximizes the logistic likelihood $\prod_{i=1}^{n_d} \mathcal{L}_i^{mis}(\alpha)$, where

$$\mathcal{L}^{mis}(\alpha) = \left\{ \prod_{m=0}^{C-1} \pi(m, \alpha) \right\} \{1 - \pi(C, \alpha)\}^{I(C \neq K+1)}.$$

The IPCW estimator is then $\hat{\beta}_w = \{\sum_{i=1}^{n_d} \delta_i Y_i / \bar{\pi}_i(K, \hat{\alpha})\} / \{\sum_{i=1}^{n_d} \delta_i / \bar{\pi}_i(K, \hat{\alpha})\}$, $\bar{\pi}(m, \alpha) \equiv \prod_{k=0}^m \pi(k, \alpha)$ and $\delta = I(\bar{R}_K = \mathbf{0}, Z = d)$. The denominator of $\hat{\beta}_w$ is a consistent estimator of the sample size n_d in arm $Z = d$. (Because of the large number of nuisance parameters α_{1m} , one might choose to estimate α_2 by conditional logistic regression, which eliminates the α_{1m} , and only then estimate the α_{1m} .)²⁹ Worked examples using this method can be found in references 7, 8, 11, 13 and 18.

3.3. Approach 3: ICE Estimator

We specify a parametric model $h(\bar{L}_k, \psi_k)$ for η_k in arm $Z = d$ where $h(\cdot, \cdot)$ is a known function. We let $\hat{\psi}_K$ be the possibly non-linear least squares estimator of ψ_K from the regression of Y on \bar{L}_K with regression function $g_K(\bar{L}_K, \psi_K)$ among subjects with $\bar{R}_K = \mathbf{0}$ and $Z = d$. Then recursively, for $K - 1, \dots, 0$, we let $\hat{\psi}_k$ be the possibly non-linear least squares estimator of ψ_k from the regression of the ‘outcome variable’ $h_{k+1}(\bar{L}_{k+1}, \hat{\psi}_{k+1})$ on \bar{L}_k with regression function $h_k(\bar{L}_k, \psi_k)$ among subjects with $\bar{R}_k = \mathbf{0}$ and $Z = d$. Then the ICE estimator is $\hat{\beta}_1 = n_d^{-1} \sum_{i=1}^{n_d} h_0(\bar{L}_{0i}, \hat{\psi}_0)$. If $h_k(\bar{L}_k, \psi_k) = \sum_{m=0}^k \sum_{j=0}^m \psi_{kmj} L_{mj}$, where $L_m \equiv (L_{m1}, \dots, L_{mJ_m})'$ has components L_{mj} , then $\hat{\beta}_1$ is the extended sweep estimator (reference 13, p. 114). By definition, in a given treatment arm, the extended sweep estimator is the maximum likelihood estimator of the mean of Y under the assumptions that (a) Y and \bar{L}_K are jointly multivariate normal with unrestricted mean and covariance matrix, and (b) the data are viewed as monotone missing at random data where we regard a subject as missing at the first time k at which the subject does not comply ($R_k = 1$). Worked examples using this method can be found in reference 13.

3.4. Approach 4: Marginal Model Estimation

Given arm $Z = d$, let $f[Y(d); \gamma_1]$ be a parametric model for the marginal distribution of $Y(d)$. In addition, let $f[L_k | \bar{L}_{k-1}, \bar{R}_{k-1} = \mathbf{0}, Y(d); \gamma_2]$ be a fully parametric model, and let $(\hat{\gamma}_1, \hat{\gamma}_2)$ maximize the likelihood $\prod_{i=1}^{n_d} \mathcal{L}_{\text{mar}, i}(\gamma_1, \gamma_2)$ where $\mathcal{L}_{\text{mar}}(\gamma_1, \gamma_2)$ equals $Q(\gamma_1, \gamma_2, Y) \equiv \prod_{m=0}^{\min(C, K)} f(L_m | \bar{L}_{m-1}, \bar{R}_{m-1} = \mathbf{0}, Y; \gamma_2) f(Y; \gamma_1)$ for subjects who remain on their assigned therapy ($C = K + 1$) and equals $\int Q(\gamma_1, \gamma_2, y) dy$ otherwise. Then $\hat{\beta}_{\text{mar}} = \int y f(y; \hat{\gamma}_1) dy$. $\hat{\beta}_{\text{mar}}$ will be the extended sweep estimator for certain model choices.

Remark: A major difficulty with estimating $E[Y(d)]$ by $\hat{\beta}_{\text{mar}}$ is that consistency requires a correctly specified model for $f[L_m | \bar{L}_{m-1}, \bar{R}_{m-1} \equiv \mathbf{0}, Y(d)]$. However, since $Y(d)$ is only observed for that subset of subjects with history $\bar{L}_{m-1}, \bar{R}_{m-1} \equiv \mathbf{0}$ who remain on treatment throughout ($\bar{R}_K = 0$), it would be difficult if not impossible for an investigator to specify a parametric model for this density believed to approximate the truth. As a contrast, the estimator $\hat{\beta}_w$ only requires that one specify a model $\pi(m; \alpha)$ for the probability of remaining on treatment at time m given \bar{L}_{m-1} among those on treatment up to m . Specification of a realistic model $\pi(m; \alpha)$ only requires that the investigator has an understanding of the behavioural process by which patients choose their treatment at m given knowledge of their past covariate history.⁶

3.5. Comparison of Methods

Table I compares the properties of the four methods with respect to a number of criteria. In addition, we include the ITT estimator $\hat{\beta}_{\text{ITT}}$, which is the average of Y_i among subjects in treatment arm $Z_i = d$, and the estimator $\hat{\beta}_{\text{sa}}$, which is the ‘sample average’ of Y_i among subjects in arm $Z_i = d$ who stayed on their assigned treatment ($\bar{R}_K = \mathbf{0}$). The estimator $\hat{\beta}_{\text{ge}}$ is discussed in Section 4.3. In rows 1–15, a ‘no’ response is preferable to a ‘yes’. Only the IPCW estimator obtains a ‘no’ on all but one criteria except in the special case in which $\hat{\beta}_{\text{g}}, \hat{\beta}_1$ and $\hat{\beta}_{\text{mar}}$ are the extended sweep estimators.

The first criterion evaluated is whether the estimator $\hat{\beta}$ requires evaluation of high-dimensional integrals. This will only be the case for $\hat{\beta}_{\text{g}}$ (and then only if L_k has continuous components), although $\hat{\beta}_{\text{mar}}$ will require numerical or Monte Carlo evaluation of one-dimensional integrals for

Table I. Estimators of $E[Y(d)]$

Row	Criterion	$\hat{\beta}_g$	$\hat{\beta}_w$	$\hat{\beta}_l$	$\hat{\beta}_{mar}^*$	$\hat{\beta}_{sa}$	$\hat{\beta}_{ITT}$	$\hat{\beta}_{ge}^*$
1	Requires evaluation of high-dimensional integrals	yes*	no	no	no	no	no	no
2	Requires modelling multivariate outcomes	yes*	no	no	no	no	no	no
3	Cannot use standard software	yes*	no	no	possibly*	no	no	no
4	Cannot correct for non-random non-compliance [§]	no	no	no	no	yes	no	no
5	Can be inconsistent under random non-compliance [§]	yes*	no	yes*	yes*	no	yes	yes
6	Can be inconsistent under full compliance	yes*	no	yes*	sometimes* [†]	no	no	no
7	Super efficient under full compliance	yes*	no	yes*	sometimes* [†]	no	no	no
8	Can be inconsistent under sharp null (1)–(2)	yes	yes	yes	yes	yes	no	yes
9	Can be inconsistent under random non-compliance and null (1)–(2)	yes	no	yes	yes	no	no	no
10	Can be inconsistent under full compliance and null (1)–(2)	yes	no	yes	sometimes* [†]	no	no	no
11	Difficult to extend to regression models	yes	no	no	no	no		no
12	Difficult to extend to non-ignorable missingness models	yes	no	yes	no	yes		no
13	Difficult to extend to repeated measures outcomes	no	no	yes	somewhat [‡]	no		
14	Based on likelihood function	yes	no	no	yes	no		
15	Requires modelling distributions about which little is known	somewhat	no	somewhat	yes	no		no
16	Efficiency	more	less	more	more	least		

* In the special case $\hat{\beta}$ is the extended sweep estimator, these entries are 'no'

[†] No if $f(y; \gamma_1)$ is normal for Y continuous and is saturated for Y discrete

[‡] Dimension of integrals that must be evaluated in likelihood function will be greater

[§] $\hat{\beta}_{mar}^*$ and $\hat{\beta}_{mar}^{**}$ have the same entries, except for row 16

|| $\hat{\beta}_{ge}^*$ uses all outcomes Y , so generally more efficient than estimators that do not. Less efficient than $\hat{\beta}_{mar}^*$ and $\hat{\beta}_{mar}^{**}$

* Applies to G-estimation of any structural model

continuous Y . The second criterion is whether the estimator requires modelling of high-dimensional multivariate mixed continuous and discrete outcomes. This will be the case for $\hat{\beta}_g$ and $\hat{\beta}_{\text{mar}}$ since the conditional distribution of L_k must be estimated. The third criterion evaluated is whether the estimators can be computed with standard software. The IPCW estimator $\hat{\beta}_w$ and the ICE estimator $\hat{\beta}_I$ are particularly easy to compute, essentially requiring only logistic regression and non-linear least squares software, respectively.

The next criteria represent important robustness criteria. All four methods are designed to correct for non-random non-compliance. It would seem perverse to recommend a method to correct for non-random non-compliance that can be inconsistent under either random non-compliance (in the sense that equations (3) and (4) and thus (5) hold) or full compliance since the sample average $\hat{\beta}_{\text{sa}}$ among compliers is consistent in these settings. However, $\hat{\beta}_g$, $\hat{\beta}_I$, $\hat{\beta}_{\text{mar}}$ and $\hat{\beta}_{\text{ITT}}$ all may be inconsistent under random non-compliance. $\hat{\beta}_g$, $\hat{\beta}_I$ and $\hat{\beta}_{\text{mar}}$ can even be inconsistent under full compliance due to model misspecification. This result is closely related to the fact that $\hat{\beta}_g$, $\hat{\beta}_I$ and $\hat{\beta}_{\text{mar}}$ can all be more efficient than a simple sample average under full compliance, when the marginal distribution of $Y = Y(d)$ is non-normal. We view this 'super efficiency' as a drawback, since the data analyst will incorrectly conclude that there is more information about $E[Y(d)]$ than is actually available in the data – the other information being supplied by the (possibly incorrect) assumptions encoded in the parametric and regression models used in computing $\hat{\beta}_g$, $\hat{\beta}_I$ and $\hat{\beta}_{\text{mar}}$.

The results described in the preceding paragraph are true under the equivalence null hypothesis (1) that $Y_i(0) = Y_i(1)$. However, under the sharp null hypothesis that both (1) and (2) hold, the ITT estimator, in contrast to any of the other estimators, is consistent for $E[Y(d)]$ under arbitrary and unspecified non-random non-compliance.

Remark: Because, as discussed in Section 2, we are assuming (2) is known to be false, this last result is of little importance to us. However, if we did not know (2) were false *a priori*, we might adopt methods that use the randomization indicator Z as an instrument, since such methods, like the ITT estimator, are consistent under (1)–(2), but, under additional assumptions, can also succeed in correcting for non-compliance if (1)–(2) are false.^{5,9,16,17,21}

Often one wishes to specify a model for the regression of $Y(d)$ on treatment arm and various pre-treatment baseline covariates, such as age, race and sex. All the approaches except the G-computation algorithm estimator can be easily extended to the regression set-up. Furthermore, the IPCW estimator and $\hat{\beta}_{\text{mar}}$ are the easiest of the estimators to extend to 'non-ignorable' non-compliance models (see Section 9). All four approaches can be extended to multivariate repeated measure outcomes except for the ICE approach, because there may be no multivariate distribution that satisfies a given iterated conditional expectations model in the multivariate case (reference 13, p. 120). $\hat{\beta}_w$ and $\hat{\beta}_I$ are semi-parametric estimators and thus are not based on a parametric likelihood function. In contrast, $\hat{\beta}_g$ and $\hat{\beta}_{\text{mar}}$ are parametric maximum likelihood estimators. Row 15 of Table I reiterates the Remark in Section 3.4.

4. EFFICIENCY INCREASING ASSUMPTIONS

In row 16 of Table I, rough efficiency comparisons are made among the estimators. Except for $\hat{\beta}_{\text{ITT}}$, all the estimators in Table I ignored all data collected on a subject after the first time C that the subject left their assigned therapy. In particular, data on the outcome Y was ignored among the non-compliers ($\bar{R}_K \neq \mathbf{0}$). Hence, if the number of subjects who remained on their assigned

treatment is quite small (say, 50 per cent or less), then all the estimators in Table I (except $\hat{\beta}_{\text{ITT}}$) will have large variability, and one might wish to consider imposing additional assumptions to decrease variability and thus, if the additional assumptions are correct, mean-squared error. In fact, all the previous estimators except for $\hat{\beta}_w$ can be made substantially more efficient by incorporating additional modelling assumptions. $\hat{\beta}_w$ cannot be made substantially more efficient because only the outcomes Y in subjects who remain on therapy contribute to the estimator.

4.1. G-computation algorithm and iterated conditional expectation estimators

$\hat{\beta}_g$ can be made more efficient by specifying parsimonious parametric models for the conditional laws of L_m and Y given past covariate and compliance history rather than specifying, as we did above, the model only for the fully compliant past histories $\bar{R}_{m-1} = \mathbf{0}$ and $\bar{R}_K = \mathbf{0}$. That is, we redefine

$$\mathcal{L}_g(\theta_1, \theta_2) = \prod_{m=0}^K f(L_m | \bar{L}_{m-1}, \bar{R}_{m-1}; \theta_1) f(Y | \bar{L}_K, \bar{R}_K; \theta_2).$$

This approach allows us to borrow information concerning the conditional laws of L_m and Y for compliant subjects with $\bar{R}_{m-1} = \mathbf{0}$ and $\bar{R}_K = \mathbf{0}$ from covariate and outcome data on subjects who have abandoned their assigned therapy.

Similarly $\hat{\beta}_I$ can be made more efficient by specifying a regression function $h_K(\bar{L}_K, \bar{R}_K, \psi_K)$ for the mean of Y given \bar{L}_K, \bar{R}_K among all subjects with $Z = d$. Specifically, let $\hat{\psi}_K$ be the non-linear least squares estimate of ψ_K based on all subjects. Then recursively for $K = 1, \dots, 0$, we let $\hat{\psi}_k$ be the non-linear least squares estimator of $h_{k+1}(\bar{L}_{k+1}, \bar{R}_{k+1} = \mathbf{0}, \psi_{k+1})$ on \bar{L}_k, \bar{R}_k with a user-specified regression function $h_k(\bar{L}_k, \bar{R}_k; \psi_k)$. Then

$$\hat{\beta}_I \equiv n_d^{-1} \sum_{i=1}^{n_d} h_0(\bar{L}_{0i}, \bar{R}_{0i} = \mathbf{0}, \hat{\psi}_0).$$

Each regression borrows strength by using the data from subjects with all possible compliance histories, rather than just from the compliers. The properties of these more efficient versions of $\hat{\beta}_g$ and $\hat{\beta}_I$ remain as described in Table I, with the exception of row 16 (efficiency).

4.2. Coarse rank-preserving structural distribution models (RPSDM) for continuous Y

Suppose Y has a continuous distribution. Then, the estimator $\hat{\beta}_{\text{mar}}$ can be made more efficient by specifying a coarse rank-preserving structural distribution model (RPSDM) linking the observed data to the possibly unobserved $Y(d)$. Specifically, the data follow a coarse rank-preserving structural distribution model in treatment arm $Z = d$, if $Y(d)$ depends on a known function of the observed outcome Y , the covariate history \bar{L}_C observed during the period $(0, C)$ in which the subjects remained on their assigned treatment, and on an unknown parameter ψ_0 . That is $Y(d) = H(\psi_0)$ with $H(\psi) \equiv h(Y, \bar{L}_C, C, \psi)$ a known function of Y, \bar{L}_C, C , and an unknown, possibly vector-valued, parameter ψ , where $H(\psi)$ satisfies the following conditions: (i) the consistency condition $h(Y, \bar{L}_C, C, \psi) = Y$ if $C = K + 1$ for all ψ , so that a subject's observed value Y equals, as it must, $Y(d) = H(\psi_0)$ when the subject remains on his assigned therapy ($C = K + 1$); (ii) $h(Y, \bar{L}_C, C, 0) = Y$ so $\psi_0 = 0$ represents the null hypothesis (2) that $Y_i = Y_i(d)$ for all i in treatment arm $Z = d$ (that is, that the treatment assigned in arm d has no effect on Y); and (iii) $h(Y, \bar{L}_C, C, \psi)$ is monotone increasing in Y .

Example: $H(\psi) = Y - (K + 1 - C)\psi$.

Remark: The coarse structural models introduced here differ from the structural models that co-workers and I have previously considered in references 3, 4, 6–12 in that the coarse structural model $H(\psi)$ does not depend on the covariate or treatment history past time C . As a consequence, coarse structural models are easier to understand mathematically than were my earlier structural models. However, as discussed in Section 7, it is not generally possible to specify a coarse structural model that incorporates the investigator’s prior knowledge or beliefs concerning the biological mechanism by which therapy affects Y . Thus, I will ultimately recommend against using coarse structural models to analyse data from trials with non-compliance. Nevertheless, I have introduced them here before introducing my usual structural models because (i) due to their simplicity, they serve as a useful pedagogic aide, and (ii) they help clarify an issue raised by several previous referees concerning the possibility of consistent goodness-of-fit tests for structural models.

The maximum likelihood estimator of $E[Y(d)]$ based on the data in arm $Z = d$ under the coarse RPSDM $H(\psi)$, and the parametric models $f[Y(d); \gamma_1]$ and $f[L_k | \bar{L}_{k-1}, \bar{R}_{k-1} = \mathbf{0}; Y(d); \gamma_2]$ is $\hat{\beta}_{\text{mar}}^* = \int yf(y; \hat{\gamma}_1) d\gamma$ where now the MLE $(\hat{\gamma}_1, \hat{\gamma}_2, \hat{\psi})$ maximizes the likelihood $\prod_{i=1}^n \mathcal{L}_{\text{mar},i}^*(\gamma_1, \gamma_2, \psi)$ where

$$\mathcal{L}_{\text{mar}}^*(\gamma_1, \gamma_2, \psi) = \prod_{m=0}^{\min(K,C)} f(L_m | \bar{L}_{m-1}, \bar{R}_{m-1} = \mathbf{0}, H(\psi); \gamma_2) f[H(\psi); \gamma_1] \{\partial H(\psi) / \partial Y\}.$$

The Jacobian term $\partial H(\psi) / \partial Y$ reflects the fact that the likelihood has been written as a function of the variable $H(\psi)$ rather than Y . However, if, as in the above example, $H(\psi)$ is linear in Y , the Jacobian is 1. Note that the estimator $\hat{\beta}_{\text{mar}}^*$ unlike the estimator $\hat{\beta}_{\text{mar}}$ uses the outcome data Y on all subjects and thus will be much more efficient. Like $\hat{\beta}_{\text{mar}}$, $\hat{\beta}_{\text{mar}}^*$ does not use any data on covariate history once subjects become non-compliant.

Remark: We could, however, gain additional efficiency by constructing an estimator $\hat{\beta}_{\text{mar}}^{**}$ that borrows information contained in the covariate data of non-compliant subjects to estimate γ_2 . Specifically, one replaces the model $f(L_m | \bar{L}_{m-1}, \bar{R}_{m-1} = \mathbf{0}, H(\psi); \gamma_2)$ by a more general parametric model $f(L_m | \bar{L}_{m-1}, \bar{R}_{m-1}, H(\psi); \gamma_2)$ and now lets $(\hat{\gamma}_1, \hat{\gamma}_2, \hat{\psi})$ maximize the product of the

$$\mathcal{L}_{\text{mar}}^{**}(\gamma_1, \gamma_2, \psi) = \prod_{m=0}^K f(L_m | \bar{L}_{m-1}, \bar{R}_{m-1}, H(\psi); \gamma_2) f[H(\psi); \gamma_1] \{\partial H(\psi) / \partial Y\}$$

over the n study subjects. Then $\hat{\beta}_{\text{mar}}^{**} = \int yf(y; \hat{\gamma}_1) d\gamma$.

With the exception of row 16 (efficiency), $\hat{\beta}_{\text{mar}}^*$ and $\hat{\beta}_{\text{mar}}^{**}$ satisfy the same properties in Table I as does $\hat{\beta}_{\text{mar}}$. Like $\hat{\beta}_{\text{mar}}$, the estimators $\hat{\beta}_{\text{mar}}^*$ and $\hat{\beta}_{\text{mar}}^{**}$ can be inconsistent for $E[Y(d)]$ under random non-compliance (that is, when equations (3) and (4) and thus (5) are true) even if the sharp null hypothesis (2) that $Y_i(d) = Y_i$ in arm $Z = d$ is true.

4.3. G-Estimation

The estimators $\hat{\beta}_{\text{mar}}^*$ and $\hat{\beta}_{\text{mar}}^{**}$, like $\hat{\beta}_{\text{mar}}$, suffer from the difficult task of needing to correctly specify a model for the covariates L_m given \bar{L}_{m-1} and (the often unobserved) $Y(d)$ among subjects on treatment throughout time m . As discussed in the Remark in Section 3.4, it is a much easier task to specify correct or nearly correct models for the conditional probability $\pi(m)$ of continued

compliance at time m given \bar{L}_m . G-estimation is an estimation method that will produce consistent estimators of both the parameter ψ_0 of a coarse RPSDM and of $E[Y(d)]$, provided the compliance model $\pi(m, \alpha)$ and the coarse RPSDM $H(\psi)$ are correctly specified. Hence, as discussed below, in the presence of random non-compliance, if we specify a coarse RPSDM $H(\psi)$, our estimate of $E[Y(d)]$ is guaranteed to be consistent under the sharp null (2) that $Y_i(d) = Y_i$ in arm $Z = d$, provided we use the method of G-estimation.

The G-estimate $\hat{\psi}_{ge}$ of ψ_0 is the value of ψ that makes the MLE of the parameter θ equal to zero when maximizing the logistic ‘likelihood’ $\prod_{i=1}^{n_d} \mathcal{L}_i^{\text{mis}^*}(\alpha, \theta)$ over (α, θ) with ψ held fixed where $\mathcal{L}^{\text{mis}^*}(\alpha, \theta) = \prod_{m=1}^{C-1} \pi(m, \alpha, \theta) \{1 - \pi(C, \alpha, \theta)\}^{I(C \neq K+1)}$ with $\pi(m, \alpha, \theta) = \alpha_{1m} + \alpha_2 h_m(\bar{L}_m) + \theta' g_m[H(\psi), \bar{L}_m]$ where θ and $g_m(\cdot, \cdot)$ are the dimension of ψ and $g_m(\cdot, \cdot)$ is a known function chosen by the investigator (for example, $g_m(H(\psi), \bar{L}_m) = H(\psi)$ if ψ is one-dimensional). We put the term ‘likelihood’ in parentheses because the ‘likelihood’ $\mathcal{L}^{\text{mis}^*}(\alpha, \theta)$ is not related to the true likelihood function for ψ , but rather is an artificial likelihood which we use as a computational ‘trick’ to obtain the G-estimate of ψ_0 . In fact, G-estimation is a semi-parametric rather than likelihood-based estimation procedure. (Estimation could also be carried out by maximizing the conditional logistic likelihood function which eliminates the α_{1m} nuisance parameters.) The choice of $g_m(\cdot, \cdot)$ which minimizes the variance of $\hat{\psi}_{ge}$ is discussed in Section 9. A nominal $(1 - \alpha)$ level confidence set for ψ is the set of ψ for which an α -level likelihood ratio, score, or Wald test based on the ‘likelihood’ $\prod_{i=1}^{n_d} \mathcal{L}_i^{\text{mis}^*}(\alpha, \theta)$ of the hypothesis $\theta = 0$ does not reject. If the compliance model (6) for $\pi(m)$ is correctly specified, then $\hat{\psi}_{ge}$ will be consistent for ψ_0 and, in large samples, the actual coverage level of a $(1 - \alpha)$ level confidence set will equal its nominal. Under random non-compliance (that is, (3) and (4) hold), the model (6) is guaranteed to be correctly specified with $\alpha_2 = 0$. Further, under the sharp null (2) that $Y_i(d) = Y_i$ in arm $Z = d$, the coarse RPSDM is guaranteed to be correctly specified with true value $\psi_0 = 0$. Thus, under the sharp null (2) and random non-compliance, the G-estimate $\hat{\beta}_{ge} = n_d^{-1} \sum_{i=1}^{n_d} H_i(\hat{\psi}_{ge})$ is guaranteed to be consistent for $E[Y(d)]$. However, $\hat{\beta}_{ge}$, in contrast to $\hat{\beta}_w$, is not *guaranteed* to be consistent for $E[Y(d)]$ under random non-compliance if the sharp null $Y_i(d) = Y_i$ is false, since consistency of $\hat{\beta}_{ge}$ also requires that the coarse RPSDM model be correctly specified. Under explainable non-random non-compliance (that is, when (3) holds but (4) does not), $\hat{\beta}_{ge}$ can be inconsistent if the compliance model (6) or the model $H(\psi)$ is misspecified. Properties of $\hat{\beta}_{ge}$ are summarized in Table 1. $\hat{\beta}_{ge}$ is always less efficient than $\hat{\beta}_{mar}^*$ when the assumptions guaranteeing the consistency of both are true.

5. COARSE STRUCTURAL NESTED DISTRIBUTION MODELS FOR CONTINUOUS Y

A serious difficulty with a coarse RPSDM is that the model incorporates the ‘rank preservation’ assumption that any two subjects with the same values of Y, C, \bar{L}_C will have the same value of $Y(d)$ even if one subject receives more treatment after C than the other. Hence it is biologically implausible that the data follow a coarse RPSDM when the sharp null hypothesis (2) is false. Nonetheless, we can show that the properties of $\hat{\beta}_{ge}$, $\hat{\beta}_{mar}^*$ and $\hat{\beta}_{mar}^{**}$ described in the previous section still hold if the data follow a coarse structural nested distribution model (SNDM), a non-rank preserving generalization of a coarse RPSDM, defined below.

To describe a coarse SNDM, let $H = h(Y, \bar{L}_C, C)$ be the (unknown) unique increasing function of Y such that H and $Y(d)$ have the same distribution given (\bar{L}_C, C) in arm $Z = d$. Since Y is assumed to have continuous conditional distributions, such a conditional quantile-quantile function always exists. We say we have rank preservation if $H = Y(d)$ with probability 1 as in

Section 4. However, even if $H \neq Y(d)$, nevertheless, in treatment arm $Z = d$, H and $Y(d)$ will have the same distribution given $R_k, \bar{R}_{k-1} = \mathbf{0}, \bar{L}_k$ for each k (reference 7, Appendix 1). Hence, the explainable non-random non-compliance assumption (3) implies that H and R_k are conditionally independent given \bar{L}_k among subjects remaining on treatment through time k . That is, in arm $Z = d$,

$$H \perp\!\!\!\perp R_k \mid \bar{R}_{k-1} = \mathbf{0}, \bar{L}_k. \quad (7)$$

We say that the data follow a coarse SNDM if $H = H(\psi_0)$ where $H(\psi) = h(Y, \bar{L}_C, C, \psi)$ is defined as in Section 4 and ψ_0 is an unknown parameter to be estimated. That is, a coarse SNDM assumes that the function $h(Y, \bar{L}_C, C)$ is known up to an unknown parameter ψ_0 . Our previous results do not require rank preservation, although they do require that we be able to specify the function $h(Y, \bar{L}_C, C)$ up to an unknown parameter ψ_0 . For example, given the compliance model (6) and the explainable non-random non-compliance assumption (3), it can be shown that the conditional independence relation (7) implies that $\hat{\psi}_{ge}$ and $\hat{\beta}_{ge}$ remain consistent for ψ_0 and $E[Y(d)]$ under a correctly specified coarse SNDM $H(\psi)$ (references 3 and 8 Appendix 1). Similarly, $\hat{\beta}_{mar}^*$ is consistent for $E[Y(d)]$ under a correct coarse SNDM $H(\psi)$ if the explainable non-random non-compliance assumption (3) holds and the models $f[Y(d); \gamma_1]$ and $f[L_k \mid \bar{L}_{k-1}, \bar{R}_{k-1} = \mathbf{0}, Y(d); \gamma_2]$ are correctly specified (reference 8, Appendix 1).

6. COARSE STRUCTURAL NESTED MEAN MODELS

6.1. Model Definition

Coarse structural nested mean models (SNMMs), in contrast with the coarse SNDMs, may be used to estimate $E[Y(d)]$ with either discrete or continuous Y . To describe these models, define $\gamma(C, \bar{L}_C) = E[Y - Y(d) \mid \bar{L}_C, C]$, to be the average difference between Y and $Y(d)$ among those first leaving therapy at C with covariate history \bar{L}_C in arm $Z = d$. Redefine H to be $Y - \gamma(C, \bar{L}_C)$, so that H and $Y(d)$ have the same mean given \bar{L}_C, C . The data follow a coarse SNMM $H(\psi)$ in arm $Z = d$ if H is known up to a finite dimensional unknown parameter ψ_0 . That is, $H = H(\psi_0)$ where ψ_0 is an unknown parameter, $H(\psi) \equiv Y - \gamma(C, \bar{L}_C, \psi)$ and $\gamma(C, \bar{L}_C, \psi)$ is a known function satisfying $\gamma(C, \bar{L}_C, \psi) = 0$ if $C = K + 1$, and $\gamma(C, \bar{L}_C, 0) = 0$ so $\psi = 0$ is the 'null value'. As an example, $\gamma(C, \bar{L}_C, \psi)$ could be $(K + 1 - C)\psi$.

Results of reference 12 imply that the conditional means of H and $Y(d)$ are equal given $\bar{L}_k, \bar{R}_{k-1} = \mathbf{0}, R_k$. It follows that if the explainable non-random non-compliance assumption (3) holds and the coarse SNMM $H(\psi)$ is correctly specified, then the mean of $H = H(\psi_0)$ does not depend on the treatment R_k received at time k given $\bar{L}_k, \bar{R}_{k-1} = \mathbf{0}$, that is

$$E[H \mid \bar{L}_k, \bar{R}_{k-1} = \mathbf{0}, R_k] = E[H \mid \bar{L}_k, \bar{R}_{k-1} = \mathbf{0}]. \quad (8)$$

It can be shown that (8) in turn implies that, if the compliance model (6) is correctly specified, $\hat{\psi}_{ge}$ and $\hat{\beta}_{ge}$ of Section 4 are consistent for ψ_0 and $E[Y(d)]$, provided $g_m(H(\psi), \bar{L}_m)$ is chosen linear in $H(\psi)$. That is, $\hat{\psi}_{ge}$ and $\hat{\beta}_{ge}$ are consistent provided $g_m(H(\psi), \bar{L}_m) = g_{1m}(\bar{L}_m) H(\psi) + g_{2m}(\bar{L}_m)$ where $g_{1m}(\bar{L}_m)$ and $g_{2m}(\bar{L}_m)$ are chosen by the data analyst. However, a confidence set for ψ_0 can no longer be obtained as the set of ψ for which likelihood ratio, Wald or score test of hypothesis $\theta = 0$ fails to reject. Appropriate confidence procedures are given reference 12, p. 2396.

The need for $g_m(H(\psi), \bar{L}_m)$ to be linear in $H(\psi)$ and the inability to use likelihood-based confidence sets is a consequence of the fact that the variable H , as defined for a SNMM, does not satisfy the restriction (7) satisfied by the variable H as defined for a SNDM. For

this same reason, the likelihood-based estimator $\hat{\beta}_{\text{mar}}^*$ is no longer consistent for $E[Y(d)]$ under a SNMM $H(\psi)$. In fact, likelihood-based inference under a SNMM will be not considered in this paper since it is rather complex, requiring use of the complex likelihood function derived in the appendix of reference 12. Note that the coarse SNDM $H(\psi) = Y - \psi(K + 1 - C)$ is also a coarse SNMM, but a coarse SNMM $H(\psi) = Y - \psi(K + 1 - C)$ need not be a coarse SNDM.

Because coarse SNMMs do not automatically impose the restriction that $E[Y(d) | \bar{L}_C, C] \geq 0$, they may not be satisfactory if Y and $Y(d)$ are non-negative, as, for example, when Y represents a Poisson or overdispersed Poisson random variable. In this setting, we can use a multiplicative coarse structural nested mean model. Specifically, redefine $\gamma(C, \bar{L}_C) = \ln\{E(Y | \bar{L}_C, C) / E(Y(d) | \bar{L}_C, C)\}$ and $H = Y \exp\{-\gamma(C, \bar{L}_C)\}$. We say the data follow a multiplicative coarse SNMM $H(\psi)$ if H is known up to a finite dimensional parameter ψ_0 , that is, $H = H(\psi)$, $H(\psi) = Y \exp\{-\gamma(C, \bar{L}_C, \psi)\}$ and $\gamma(C, \bar{L}_C, \psi)$ is as defined above. $\hat{\psi}_{\text{ge}}$ and $\hat{\beta}_{\text{ge}}$ remain consistent estimators of ψ_0 and $E[Y(d)]$ under a multiplicative coarse SNMM when (6) and (3) hold.¹²

Neither a coarse SNMM nor a multiplicative coarse SNMM automatically imposes the true restriction $0 \leq E[Y(d) | C, \bar{L}_C] \leq 1$ when Y and $Y(d)$ are Bernoulli random variables. Unfortunately, there is no simple method to estimate the parameter ψ_0 of a 'logistic' coarse SNMM that imposes the restriction $\gamma(C, \bar{L}_C, \psi_0) = \text{logit } E[Y | \bar{L}_C, C] / \text{logit } E[Y(d) | \bar{L}_C, C]$. However, if $E[Y(d) | C, \bar{L}_C]$ is always small (that is, under the rare disease assumption), one can approximate a logistic coarse SNMM with a coarse multiplicative SNMM. With Y Bernoulli, we might also consider forgoing structural nested models and instead use the iterated conditional expectations estimator $\hat{\beta}_1$ with $h_k(\bar{L}_k, \bar{R}_k, \psi_k)$ given by the logistic function $[1 + \exp\{-\psi'_k W_k\}]^{-1}$ with W_k a vector function of (\bar{L}_k, \bar{R}_k) . This choice will ensure that estimated probabilities remain between zero and 1.

6.2. Comparison of coarse SNMMs with coarse SNDMs for continuous Y

For continuous Y , coarse SNDMs have two advantages over coarse SNMMs. First, coarse SNMMs only allow one to estimate the mean of $Y(d)$ without further model assumptions, while coarse SNDMs allows one to estimate the entire distribution of $Y(d)$ which will be important if the analyst wishes to compare aspects of the distributions of $Y(1)$ and $Y(0)$ other than their means. Specifically, in arm $Z = d$, the empirical distribution of the $H_i(\hat{\psi}_{\text{ge}})$ is a $n^{1/2}$ -consistent estimator of the distribution of $Y(d)$ when the $H_i(\hat{\psi}_{\text{ge}})$ are based on the fit of correctly specified SNDMs but not when based on the fit of an SNMM. Hence an investigator who, in the absence of non-compliance, would have compared the distributions of Y in treatment arm $Z = 1$ with that in treatment arm $Z = 0$ by using a standard two-sample test (for example, the Wilcoxon test or logrank test) can continue to use the same statistic numerator test to compare the distribution of the $H_i(\hat{\psi}_{\text{ge}})$ in the two arms based on a correctly specified SNDM (although a correction to the variance of the numerator will be required to account for the fact that, in each treatment arm, there is additional variability due to $\hat{\psi}_{\text{ge}}$ being an estimate of the true ψ_0). A worked example is provided in reference 11.

Of course this advantage of coarse SNDMs compared to coarse SNMMs comes at a price; it is much more difficult to correctly specify an SNDM than an SNMM since an SNDM model requires one specify a correct model for the conditional quantile-quantile plot of Y versus $Y(d)$ given C, \bar{L}_C (rather than just for the difference in conditional means).

The second advantage of an SNDM over an SNMM is that the function $g_m(H(\psi), \bar{L}_m)$ added to the logistic model $\pi(m, \alpha)$ in conducting G-estimation need not be linear in $H(\psi)$, so one may bound the influence of any single observation on $\hat{\psi}_{ge}$ by choosing a function that down-weights large or small values of $H(\psi)$ such as $g_m(H(\psi), \bar{L}_m) = \text{sign}[H(\psi)]\min\{|H(\psi)|, c\}$ for some positive constant c . That is, one can use ‘bounded influence’ G-estimation to ensure robustness.³⁰

7. STRUCTURAL NESTED MODELS INCORPORATING *A PRIORI* BIOLOGICAL KNOWLEDGE

An apparent benefit of the G-estimate $\hat{\beta}_{ge}$ or likelihood-based estimate $\hat{\beta}_{mar}^*$ of a coarse SNDM (or coarse SNMM) is that the estimators do not use data on treatment history R_k or covariate history L_k after the first time C that a subject leaves his/her assigned treatment, thus simplifying the analysis. However, this apparent benefit only accrues to the data analyst once a coarse SNDM or coarse SNMM has been specified. Specification of a coarse structural nested model (mean or distribution), like all models, must be based on the investigator’s prior knowledge and beliefs. As we now show, it is not generally possible to specify a coarse structural nested model (SNM) that adequately incorporates an investigator’s knowledge and beliefs. In contrast, we will show that specification of the SNMs previously described in references 3, 4, 6–12 (that do depend on treatment and covariate data after time C) allow one to incorporate such prior information. Henceforth, when we refer to a SNM without the appellation ‘coarse’, we refer to the SNMs previously described in the above references.

To underscore the difficulties in specifying a coarse SNM, consider a subset of subjects with $C = K - 4$ and a particular covariate history \bar{L}_C . Then one’s opinions about how to model $\gamma(C, \bar{L}_C) \equiv E[Y - Y(d) | C, \bar{L}_C]$ when $C = K - 4$ would certainly depend on the fraction of subjects with $C = K - 4$ remaining off therapy at times $K - 2$, $K - 1$, and K if one knew *a priori* that treatment received at each time affected Y . However, one’s opinion would not depend on these fractions if it were known *a priori* that there was a biological latent period of three time periods (that is, treatment received at times $K - 2$ or later could not influence Y measured at $K + 1$). It is not obvious how one would modify one’s specification of the coarse SNMM model $\gamma(C, \bar{L}_C, \psi)$ (which does not depend on the data R_{K-2} , R_{K-1} , or R_K) in order to incorporate *a priori* knowledge of a latent period. We now describe how the prior knowledge of a biological latent period can be directly incorporated in the SNMMs of reference 12.

7.1. SNMMs

Let $Y_{(\bar{R}_k, 0)}$ be the possibly counterfactual outcome of a subject in arm $Z = d$ who follows his observed treatment history \bar{R}_k until time $k + 1$ and then stays on his assigned therapy thereafter. Then $Y_{(0)}$ is, by definition, the possibly counterfactual outcome if a subject remains on his assigned therapy throughout. That is, $Y_{(0)} = Y(d)$. Also, by definition, $Y_{(\bar{R}_k, 0)} \equiv Y_{(\bar{R}_k)}$ is the outcome of a subject who follows his observed treatment history until the time $K + 1$ at which Y is measured. Thus, by definition, $Y_{(\bar{R}_k, 0)} = Y$. Let $\gamma(k, \bar{L}_k, \bar{R}_k) = E[Y_{(\bar{R}_k, 0)} - Y_{(\bar{R}_{k-1}, 0)} | \bar{L}_k, \bar{R}_k]$ be the difference, among subjects with observed history \bar{L}_k, \bar{R}_k , of the mean of Y that would have been observed if the subjects, contrary to fact, had remained on therapy after time $k + 1$ and the mean of Y had subjects remained on therapy after k . Hence, if the observed treatment R_k is the assigned treatment (that is, $R_k = 0$), then $\gamma(k, \bar{L}_k, \bar{R}_k) = 0$.

Redefine $H \equiv Y - \sum_{k=0}^K \gamma(k, \bar{L}_k, \bar{R}_k)$. Note H is a function of all the data $(\bar{L}_K, \bar{R}_K, Y)$. H is an estimate of $Y(d)$ since it ‘removes’ the effect of the observed treatment at times $K, K - 1$, and so on until time 0 and replaces these effects with those of the assigned treatment. In reference 12, I show that under the assumption of explained non-random non-compliance (3), the conditional mean of H does not depend on the observed treatment R_k , that is, equation (8) holds. We say the data follow an SNMM $H(\psi)$ if H is known up to an unknown parameter. That is, $H = H(\psi_0)$ and $\gamma(k, \bar{L}_k, \bar{R}_k) = \gamma(k, \bar{L}_k, \bar{R}_k, \psi_0)$, where $H(\psi) = Y - \sum_{k=0}^K \gamma(k, \bar{L}_k, \bar{R}_k, \psi)$, ψ_0 is an unknown parameter, and the known function $\gamma(k, \bar{L}_k, \bar{R}_k, \psi)$ is zero if either $R_k = 0$ or $\psi = 0$. The estimators $\hat{\psi}_{ge}$ and $\hat{\beta}_{ge}$ of Section 4 with $g_m(H(\psi), \bar{L}_m)$ chosen linear in $H(\psi)$ are consistent for ψ_0 and $E[Y(d)]$ when (3) and (6) hold and the SNMM $H(\psi)$ is correctly specified.

A simple example of a SNMM without a biological latent periods has $\gamma(k, \bar{L}_k, \bar{R}_k, \psi) = \psi R_k$. To incorporate a biological latent period of three time periods, we simply modify this model to $\gamma(k, \bar{L}_k, \bar{R}_k, \psi) = \psi R_k I(k < K - 2)$ so that $\gamma(k, \bar{L}_k, \bar{R}_k, \psi) = 0$ for all ψ if $k = K - 2, K - 1$, or K . Multiplicative SNMMs can be defined in an analogous fashion to multiplicative coarse SNMMs.¹²

7.2. SNDMs

We now describe how, for continuous Y , prior information such as a biological latent period can also be incorporated directly into structural nested distribution models (SNDMs) described in Section A2.16 of reference 20. To define an SNDM, let $\Gamma_k \equiv \gamma_k(Y_{(\bar{R}_k, 0)}, \bar{L}_k, \bar{R}_k)$ be the unique (unknown) function such that Γ_k and $Y_{(\bar{R}_k, 0)}$ have the same distribution given \bar{L}_k, \bar{R}_k . If, rather than only having the same conditional distributions, Γ_k equals $Y_{(\bar{R}_k, 0)}$ with probability 1, we say we have (local) rank preservation which would usually be biologically implausible. Note however that when $R_k = 0$, $\Gamma_k = Y_{(\bar{R}_k, 0)} = Y_{(\bar{R}_{k-1}, 0)}$ with probability one by the very definition of $Y_{(\bar{R}_k, 0)}$. Also note $\gamma_K(Y, \bar{L}_K, \bar{R}_K) \equiv \Gamma_K$ since, by definition, $Y = Y_{(\bar{R}_K, 0)}$.

Example: $\Gamma_k = Y_{(\bar{R}_k, 0)} - \psi_0 R_k$.

Let $H_K = \Gamma_K$. For $k = K - 1, \dots, 0$, define H_k recursively by $H_k = \gamma_k(H_{k+1}, \bar{L}_k, \bar{R}_k)$. Redefine H to be H_0 . Note under (local) rank preservation, $H_k = \Gamma_k = Y_{(\bar{R}_k, 0)}$ and thus $H = Y_{(0)} \equiv Y(d)$. Hence, under the explainable non-random non-compliance assumption (3), (7) still holds. In fact, I show in Appendix 1 of reference 8 that, even without local rank preservation, (3) implies (7) and that H and $Y(d)$ have the same marginal distribution.

We say that the data follow an SNDM if H is known up to an unknown parameter ψ_0 , that is, $H = H(\psi_0)$. This is equivalent to saying that $\gamma_k(y, \bar{L}_k, \bar{R}_k)$ is known up to an unknown parameter ψ_0 . That is, (i) $\gamma_k(y, \bar{L}_k, \bar{R}_k) = \gamma_k(y, \bar{L}_k, \bar{R}_k, \psi_0)$ where $\gamma_k(y, \bar{L}_k, \bar{R}_k, \psi)$ is a known function that takes the value y when $R_k = 0$ or $\psi = 0$, and $\gamma_k(y, \bar{L}_k, \bar{R}_k, \psi)$ is increasing in y , and (ii) $H(\psi)$ is defined like H except with $\gamma_k(y, \bar{L}_k, \bar{R}_k, \psi)$ replacing the unknown $\gamma_k(y, \bar{L}_k, \bar{R}_k)$.

When (3) and (6) hold and our SNDM $H(\psi)$ is correctly specified, it can be shown that the conditional independence relation (7) implies that $\hat{\psi}_{ge}$ and $\hat{\beta}_{ge}$ are consistent for ψ_0 and $E[Y(d)]$ (reference 8, Appendix 1). Similarly, $\hat{\beta}_{mar}^*$ is consistent for $E[Y(d)]$ under a correct SNDM $H(\psi)$ if the explainable non-random non-compliance assumption (3) holds and the models $f[Y(d); \gamma_1]$ and $f[L_k | \bar{L}_{k-1}, \bar{R}_{k-1} = \mathbf{0}, Y(d); \gamma_2]$ are correctly specified (reference 8, Appendix 1).

Example: If $\gamma_k(y, \bar{L}_k, \bar{R}_k, \psi) = y - \psi R_k$ then $H(\psi) = Y - \sum_{k=0}^K \psi R_k$ and our SNDM is an SNMM as well. In contrast, the SNMM example of Section 7.1 need not be an SNDM. If we impose the restriction that $\gamma_k(y, \bar{L}_k, \bar{R}_k, \psi) = y$ when $k = K - 2, K - 1$, or K (for example, $\gamma_k(y, \bar{L}_k, \bar{R}_k, \psi) = Y - R_k \psi I(k < K - 2)$), we have imposed a three period biological latent period. The analyses of both randomized and observational studies using SNDMs are described in reference 8, Section 5.

7.3. Goodness-of-Fit in Structural Nested Models

Given the explainable non-random non-compliance assumption (3), the functions $h(Y, \bar{L}_C, C)$ and $\gamma(C, \bar{L}_C) \equiv E(Y|C, \bar{L}_C) - E[Y(d)|C, \bar{L}_C]$ that are modelled in coarse SNDMs and coarse SNMMs, respectively, are identified (that is, are functions of the joint distribution of the observed data). For example, in arm $Z = d$, $E[Y(d)|C, \bar{L}_C]$ is the conditional IPCW estimand

$$E \left[I(\bar{R}_K = \mathbf{0}) Y / \left\{ \prod_{m=c+1}^{m=K} \pi(m) \right\} \middle| > c, \bar{L}_c \right]$$

evaluated at $C = c$ so $\gamma(C, \bar{L}_C)$ is identified. In contrast, the functions $\gamma_k(y, \bar{L}_k, \bar{R}_k)$ and $\gamma(k, \bar{L}_k, \bar{R}_k)$ that are modelled in SNDMs and SNMMs are not identified (that is, there are many such functions consistent with the distribution of the observed data). Thus in principle there exists consistent tests of the hypothesis that a coarse SNDM model or a coarse SNMM model is correctly specified but not of the hypothesis that a SNDM or an SNMM model is correctly specified. That is, in contrast to SNMs, if coarse SNMs are misspecified, then, given a sufficiently large sample size, we can construct a test that will reject the hypothesis of correct specification with probability approaching 1. For example, the test that rejects when a non-parametric (for example, kernel smoothed) estimate of $E(Y|\bar{L}_C, C)$ minus $\gamma(C, \bar{L}_C, \hat{\psi}_{ge})$ differs from a non-parametric estimate of the conditional ICE estimand by more than a fixed constant, say, 0.001, will be consistent test.

We would have identification for SNMs and thus the ability to construct consistent tests if the explainable non-random non-compliance assumption (3) were strengthened to

$$R_k \perp\!\!\!\perp Y_{(\bar{r}_k)} | \bar{R}_{k-1} = \bar{r}_{k-1}, \bar{L}_k \quad (9)$$

for all treatment histories \bar{r}_K . Equation (9) would hold if the treatment R_k were assigned at random at time k given the past (that is, among subjects with the same past treatment and covariate history), while (3) would hold if treatment R_k were assigned at random given the past only for subjects yet to leave their assigned treatment.

Remark: This follows from the fact that, in reference 2, I show that the distribution of $Y_{(\bar{r}_k)}$ conditional on $\bar{L}_k, \bar{R}_{k-1} = \bar{r}_{k-1}$ is identified under (9) for each treatment history \bar{r}_K . The functions $\gamma_k(y, \bar{L}_k, \bar{R}_k)$ and $\gamma(k, \bar{L}_k, \bar{R}_k)$ are functions of these conditional distributions. Indeed, assumption (9) would allow us to estimate the distribution of the counterfactual outcome $Y_{(\bar{r}_k)}$ under treatment regimes \bar{r}_K other than those assigned.

In some trials, subjects who deviate from their assigned treatment often also miss clinic visits. In that case, data on the factors that induce them to either return to their assigned therapy or to continue off that therapy are not available for data analysis and thus cannot be recorded in \bar{L}_k ; hence, the assumption (9) would be much less reliable than the weaker assumption (3).

Based on the above discussion, one might suppose that, in conflict with our earlier argument when we are only willing to assume (3), it might be more prudent to use coarse SNMs than SNMs because of the availability, at least in principle, of consistent goodness-of-fit tests. This argument is not correct for two reasons, one practical and one theoretical. As a practical matter, we essentially never have a sufficiently large sample size to have enough power to detect most forms of misspecification of a coarse SNM. Rather, in practice, we usually are satisfied with a goodness-of-fit test based on nesting our structural model in a larger structural model (that is, one with more parameters) and testing whether the additional parameters are compatible with zero.

From a theoretical point of view, suppose our ultimate goal remains to consistently estimate $E[Y(d)]$. Thus we would not care if we have misspecified an SNM if our estimate of $E[Y(d)]$ remains consistent under such misspecification. Now, even if an SNMM is misspecified (in the sense that there is no value ψ_0 of ψ for which the function $\gamma(k, \bar{L}_k, \bar{R}_k)$ equals the function $\gamma(k, \bar{L}_k, \bar{R}_k, \psi_0)$), nonetheless, if the hypothesis that there exists a value ψ^* of ψ such that $E[Y - \sum_{k=C}^K \gamma(k, \bar{L}_k, \bar{R}_k, \psi^*) | C, \bar{L}_C]$ is equal to the conditional IPCW estimand for all C, \bar{L}_C is true, then the G-estimate, $\hat{\psi}_{ge}$ will converge to ψ^* and $\hat{\beta}_{ge}$ will be consistent for $E[Y(d)]$ provided the non-compliance model (6) is correctly specified. However, since the conditional IPCW estimand is identified, it is possible to construct a consistent goodness-of-fit test of the above hypothesis that such a ψ^* exists. Thus, for estimating $E[Y(d)]$, the benefits of using structural nested models in comparison to coarse structural nested models described above are not countermanded by the inability to construct consistent goodness-of-fit tests of SNMs under assumption (3).

7.4. SNM versus Marginal structural models, G-computation algorithm and iterated conditional expectations

Suppose, as discussed in Section 4, it is necessary for reasons of efficiency to use data on subjects after the time C that they first leave their assigned therapy. Now, the G-computation algorithm estimator of Section 4.1 depends on parameters $\theta = (\theta_1, \theta_2)'$ indexing parametric models for the law of L_m given \bar{L}_{m-1} and \bar{R}_{m-1} and for the law of Y given \bar{L}_K and \bar{R}_K . As with a coarse SNM, it is not possible to use these parametric models to incorporate prior biological knowledge. For example, knowledge of a biological latent period of three time intervals does not imply any simple functional constraints (such as subvector of θ is equal to 0) on the parameter θ . Indeed, the situation is worse than for coarse SNMs. Even the null hypothesis (2) that $Y_i = Y_i(d)$ fails to imply a simple functional constraint for the parameters θ . In contrast, the null hypothesis (2) implies the entire parameter vector ψ of a coarse SNM is zero. For these reasons, estimation of $Y(d)$ based on SNMs is often to be preferred to estimation based on coarse SNMs or the G-computation algorithm estimator.

However, if (7.1) holds, by results of Robins,^{41,42} we can directly incorporate *a priori* biological knowledge by specifying a (non-nested) marginal structural mean model $E[Y_{(\bar{r}_k)}] = h(\bar{r}_k, \psi_0)$, $h(\bar{r}_k, \psi)$ a known function (e.g., logistic if Y is dichotomous) and then obtain a $n^{\frac{1}{2}}$ -consistent estimators $\hat{\psi}$ of ψ_0 and $h(0, \hat{\psi})$ of $E[Y_{(0)}]$ by solving

$$0 = \sum_i \{Y_i - h(\bar{R}_{K_i}, \psi)\} t(\bar{R}_{K_i}) / W_i(\bar{R}_{K_i}; \hat{\alpha})$$

where

- (i) $t(\cdot)$ is a vector-valued function of dimension of ψ chosen by the analyst,
- (ii) $\hat{\alpha}$ maximizes $\Pi_i W_i(\bar{R}_{K_i}; \alpha)$,
- (iii) $W(\bar{r}_K; \alpha) \equiv \prod_{m=1}^K \rho_m(\bar{r}_{m-1}, \bar{L}_m; \alpha)^{R_m} \{1 - \rho_m(\bar{r}_{m-1}, \bar{L}_m; \alpha)\}^{1-R_m}$, and
- (iv) $\rho_m(\bar{r}_{m-1}, \bar{L}_m; \alpha)$ is a correct model for $pr[R_m = 1 | \bar{R}_{m-1} = \bar{r}_{m-1}, \bar{L}_m]$. Note $\hat{\psi}$ solves a non-linear least squares generalized estimating equation (GEE) weighted by an estimate $\{W_i(\bar{R}_{K_i}; \hat{\alpha})\}^{-1}$ of the inverse of the subjects' probability of having their own observed treatment history and, unlike SNMMs, always works for dichotomous Y_i .

Furthermore, both prior biological knowledge of a latent period and the null (2) can imply simple functional constraints on the parameters ψ of the regression models of Section 4.1 used to construct the iterated conditional expectations estimator $\hat{\beta}_I$. Specifically, let \min denote the minimal latent period. Then for $k \in \{K, K - 1, \dots, m\}$, $m \equiv K + 1 - \min$, in the notation of Section 4.1, $E[\eta_{k+1} | \bar{L}_k, \bar{R}_{k-1} \equiv 0, R_k] \equiv E[H_{k+1} | \bar{L}_k, \bar{R}_{k-1} \equiv 0, R_k]$ does not depend on R_k , where $H_{k+1} \equiv h_{k+1}(\bar{L}_{k+1}, \bar{R}_{k+1} \equiv 0, \psi_{k+1})$ for $k \leq K$ and $\eta_{K+1} \equiv H_{K+1} \equiv Y$. Hence, for $k \geq m$, if, as an example, $h_k(\bar{L}_k, \bar{R}_k, \psi_k)$ was parametrized as $h_{k1}(\bar{L}_k, \bar{R}_k, \psi_{k1}) I(\bar{R}_{k-1} \neq 0) + h_{k2}(\bar{L}_k, \psi_{k2}) I(R_{k-1} \equiv 0) + \psi_{k3} R_k h_{k3}(\bar{L}_k) I(\bar{R}_{k-1} \equiv 0)$ then, we conclude $\psi_{k3} = 0$.

8. CONTINUOUS TIME SNMs

There are two difficulties with the SNMs of Section 7, both of which can be solved by defining SNMs in continuous time. First, the meaning of the parameter ψ depends on the time between measurements. For example, the meaning of ψ depends on whether time m is a day versus a month later than time $m - 1$. Thus it would be advantageous to have the parameter ψ defined in terms of the instantaneous effect of a particular treatment rate. Second, if the covariate process L and/or the treatment process R can (randomly) jump in continuous time rather than just at the prespecified times $1, 2, \dots, K$, the SNMs of Section 7 cannot be used.

To extend SNMs to continuous time, we shall assume that Y is still measured at time $K + 1$ but now a subject's covariate process $\bar{L}(t) = \{L(u); 0 \leq u \leq t\}$ and treatment process $\bar{R}(t) = \{R(u); 0 \leq u \leq t\}$ are generated by a marked point process where, for example, $L(u)$ is recorded treatment at time u , that is: (i) $L(t)$ and $R(t)$ have sample paths that are CADLAG step functions, that is, they are right-continuous with left-hand limits; (ii) the $L(t)$ and $R(t)$ process do not jump simultaneously; and (iii) the total number of jumps K^* of the joint $(\bar{R}(t), \bar{L}(t))$ processes in $[0, K + 1]$ is random and finite, occurring at random times T_1, \dots, T_{K^*} . We choose this restricted class of sample paths because their statistical properties are well understood.³¹ As discussed further below, in this section we no longer assume that $R(t)$ is dichotomous. Occasionally, we will also consider the possibility that both the number of jumps K^* and their times of occurrence are fixed rather than random (as earlier in the paper). Now, given a treatment history $\bar{r} \equiv \bar{r}(K + 1) = \{r(u); 0 \leq u \leq K + 1\}$ that is a CADLAG step function on $[0, K + 1]$, let $Y_{\bar{r}}$ be the counterfactual value of Y under treatment \bar{r} and let $Y_{(\bar{r}(t), 0)}$ be the counterfactual value of Y under the treatment history $\bar{r}^*(K + 1)$ where $\bar{r}^*(u) = r(u)$ for $u \leq t$ and $\bar{r}^*(u) = 0$ otherwise. Similarly, let $Y_{(\bar{r}(t-), 0)}$ be the counterfactual value of Y under the treatment history $\bar{r}^*(K + 1)$ with $\bar{r}^*(u) = r(u)$ for $u < t$, $\bar{r}^*(u) = 0$ otherwise.

We assume $Y_{\bar{r}}$ obeys the following natural consistency assumption that essentially asserts that the future cannot affect the past.

Consistency Assumption 1a: Given \bar{r} and \bar{r}^* such that $\bar{r}(u) = \bar{r}^*(u)$, $Y_{(\bar{r}(u^-), 0)} = Y_{(\bar{r}^*(u^-), 0)}$.

The following consistency assumption links the counterfactual variables Y_r to the observable variables.

Consistency Assumption 1b: $Y = Y_{(\bar{R}(K+1^-), 0)}$ w.p.1.

The following assumption asserts that an instantaneously brief bit of treatment has a negligible effect on the distribution of Y .

Assumption 2: $\text{pr}[Y_{(\bar{R}(u^-), 0)} > y | \bar{L}(t), \bar{R}(t)]$ is continuous as a function of u .

Remark: More elegantly, we would capture the fact that a brief bit of treatment has a negligible effect on Y by the following assumption which implies Assumption 2. Given the square integrable function $r(t)$, $t \in [0, K + 1]$, let $S[r(\cdot), y, t] \equiv \text{pr}[Y_{(\bar{r}(K+1))} > y | \bar{L}(t), \bar{R}(t)]$ so $S[\cdot, \cdot, \cdot]$ maps $L_2[0, K + 1] \times R^1 \times [0, K + 1]$ into R^1 where $L_2[0, K + 1]$ are the set of square integrable functions on $[0, K + 1]$. Our assumption is that $S[r(\cdot), y, t]$ is L_2 -continuous in $r(\cdot)$. That is, for all ε there exists a δ such that if, for a given $r_1(\cdot)$ and $r_2(\cdot)$, the L_2 distance between $r_1(x)$ and $r_2(x)$, $[\int_0^{K+1} [r_1(x) - r_2(x)]^2 dx]^{1/2}$, is less than δ , then the absolute value of the difference between the conditional survival curves at y of $Y_{(\bar{r}_1(K+1))}$ and $Y_{(\bar{r}_2(K+1))}$, $|S(r_1(\cdot), y, t) - S(r_2(\cdot), y, t)|$, is less than ε . Here $\delta = \delta(y, t)$ may depend on (y, t) .

8.1. Continuous Time SNMMs

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

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

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

$V(t, h)$ and $D(t)$ will be discontinuous in t at the jump times T_m because of the abrupt change in the conditioning event defining $V(t, h)$ at $t = T_m$. $D(t) dt$ is the effect of a last blip of observed treatment $R(t)$ sustained for 'instantaneous' time dt on the mean of Y compared to that of the assigned treatment.

Define $H(t) = Y - \int_t^{K+1} D(t) dt$ and redefine H to be $H(0)$. In the Appendix we prove the following theorem for each arm $Z = d$.

Theorem 8.1: $E[H(t) | \bar{L}(t), \bar{R}(t)] = E[Y_{(\bar{R}(t^-), 0)} | \bar{L}(t), \bar{R}(t)]$. In particular, $E(H) = E[Y_{(0)}]$.

We say the data follow a continuous-time SNMM $D(t, \psi)$ if $D(t) \equiv d(t, \bar{L}(t), \bar{R}(t))$ equals $D(t, \psi_0) \equiv d(t, \bar{L}(t), \bar{R}(t), \psi_0)$ where ψ_0 is a unknown parameter to be estimated and $D(t, \psi)$ is a known function continuous in t on $[T_m, T_{m+1})$ satisfying $D(t, \psi) = 0$ if $\psi = 0$ or $R(t) = 0$.

$D(t)$ is the instantaneous version of the function $\gamma(k, \bar{L}_k, \bar{A}_k)$ of Section 7. Indeed if, as earlier, the $L(t)$ process can jump only at non-random times $0, 1, \dots, K$ and the $R(t)$ process can jump only at times $0^+, \dots, K^+$, then $\gamma(k, \bar{L}_k, \bar{R}_k) = H(k) - H(k + 1) = \int_k^{k+1} D(t) dt$. Furthermore, H_k , as

defined in Section 7.1, equals $H(k)$ as defined here. In this setting, a continuous SNMM $D(t, \psi)$ induces an SNMM $\gamma(k, \bar{L}_k, \bar{R}_k, \psi)$ where $\gamma(k, \bar{L}_k, \bar{R}_k, \psi) \equiv H(k, \psi) - H(k + 1, \psi)$ with

$$H(t, \psi) \equiv Y - \int_t^{K+1} D(t, \psi) dt.$$

Hence $H(\psi)$ of Section 7.1 is given by $H(0, \psi)$. Given that (3) and (6) are true, ψ and $E[Y(d)]$ can then be estimated by G-estimation as described in Section 7.1.

Suppose again that the treatment process may jump at random times. Then the continuous time generalization of the assumption of explainable non-random non-compliance (3) in arm $Z = d$ is the assumption that $Y(d) \equiv Y_{(0)}$ is independent of the treatment $R(t)$ received at t given covariate history $\bar{L}(t^-)$ among subjects who have previously remained on their assigned therapy, that is, informally,

$$Y_{(0)} \perp\!\!\!\perp R(t) \mid \bar{R}(t^-) \equiv \mathbf{0}, \bar{L}(t^-). \quad (10)$$

When $R(t)$ is dichotomous, assumption (10) is equivalent to the formal assumption that the conditional hazard $\lambda_C(t \mid \bar{L}(t^-), Y_{(0)})$ of the random variable C which records the time a subject first left his assigned therapy does not depend on $Y_{(0)}$, that is

$$\lambda_C(t \mid \bar{L}(t^-), Y_{(0)}) = \lambda_C(t \mid \bar{L}(t^-)). \quad (11)$$

Given a correctly specified Cox model for C , that is

$$\lambda_C(t \mid \bar{L}(t^-)) = \lambda_0(t) \exp[\alpha' W(t)] \quad (12)$$

where $W(t)$ is a vector function of $\bar{L}(t^-)$, α is an unknown vector parameter, and $\lambda_0(t)$ is an unrestricted baseline hazard function, we obtain a G-estimate of the parameter ψ of the continuous time SNMM $D(t, \psi)$ by adding the term $\theta g(H(\psi), \bar{L}(t^-))$ to model (12) where $H(\psi) = H(0, \psi)$, $g(\cdot, \cdot)$ is a known function chosen by the investigator. Specifically, the G-estimate $\hat{\psi}_{ge}$ is the value of ψ for which the Cox partial likelihood estimator of θ in the expanded model is zero. Then, it can be shown that $\hat{\psi}_{ge}$ and $\hat{\beta}_{ge} \equiv n^{-1} \sum_i H_i(\hat{\psi})$ will be consistent for ψ_0 and $E[Y(d)]$ provided (10) is true, Cox model (12) is correctly specified, and $g\{H(\psi), \bar{L}(t^-)\}$ is linear in $H(\psi)$, that is, $g\{H(\psi), \bar{L}(t^-)\} = g_1\{\bar{L}(t^-)\}H(\psi) + g_2\{\bar{L}(t^-)\}$. A consistent estimator of the asymptotic variance of $\hat{\psi}_{ge}$ can be derived using methods in reference 8, Appendix 4, and reference 12, p. 2396.

If $R(t)$ is not a dichotomous random variable, then we can obtain more efficient estimators of ψ_0 and $E[Y(d)]$. As an example, suppose not all subjects in treatment arm $Z = d$ take their assigned dosage of therapy, some individuals taking more and some less, and let $R(t)$ be the difference between the dosage rate of therapy at t and the assigned dosage rate, so, as before, $R(t) = 0$ for subjects on their assigned therapy and $R(C) \neq 0$. Now let $G^*(t) = g^*(R(t), \bar{L}(t^-))$ be an arbitrary function chosen by the investigator, for example, $G^*(t) = R(t)$. Let $h_{g^*}[\bar{L}(t^-), \alpha]$ be model for the conditional mean of $G^*(t)$ given $\bar{L}(t^-)$ among subjects who first left their assigned therapy at t (that is, $C = t$), that is, $E[G^*(C) \mid C, \bar{L}(C^-)] = h_{g^*}[\bar{L}(C^-), \alpha]$. Let $\hat{\psi}_{ge}^*$ be the value of ψ for which the (possibly non-linear) least squares estimate of θ is 0 when regressing the variable $G_i^*(C_i)$ on $\bar{L}_i(C_i)$ based on the regression function $h_{g^*}[\bar{L}_i(C_i); \alpha] + \theta H_i(\psi)$ among subjects i who eventually leave treatment, that is, $C_i < K + 1$. Variance estimation can be obtained as in references 8 and 12. Finally, report the inverse (estimated) variance weighted average $\hat{\psi}_{ge, final}$ of $\hat{\psi}_{ge}$ and $\hat{\psi}_{ge}^*$. Both $\hat{\psi}_{ge, final}$ and $\hat{\beta}_{ge, final} = n^{-1} \sum_i H_i(\hat{\psi}_{ge, final})$ will be consistent if the explainable

non-random non-compliance assumption (10), the Cox model (12), the regression model $h_{g^*}[\bar{L}(t); \alpha]$, and the continuous time SNMM $D(t, \psi)$ are all correctly specified. The variance of $\hat{\psi}_{ge, final}$ will be strictly less than that of either $\hat{\psi}_{ge}$ or $\hat{\psi}_{ge}^*$ since the latter two estimators are uncorrelated. This approach is robust in the sense that it does not require that we specify a fully parametric model $f[R(t) | \bar{L}(t^-), C = t]$ for $R(t)$. A less robust approach that does require we specify a fully parametric model is discussed in reference 5. The approach just described can also be used in the presence of non-dichotomous R_k with non-continuous time coarse SNMMs and SNMMs.

8.2. Continuous time SNDM

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

$$\text{pr}[Y_{(\bar{R}(t+h^-, 0)} > Q(y, t, h) | \bar{L}(t), \bar{R}(t)] = \text{pr}[Y_{(\bar{R}(t^-, 0)} > y | \bar{L}(t), \bar{R}(t)] \quad (13)$$

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

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

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

Remark: It is important to note that we do not have to assume $Y_{(\bar{R}(t), 0)}$ is differentiable in t for $t \neq T_m$. This is scientifically important since $Y_{(\bar{R}(t), 0)}$ will be discontinuous at t if Y is a continuous measure of liver function, $R(\cdot)$ denotes exposure to a carcinogenic chemotherapeutic agent, a single molecule of which can initiate liver cancer. However, Assumptions 2 and 4 remain reasonable, since the probability of liver cancer being initiated in $[t, t + \delta t())$ is small. Assumptions 2 and 4 would be inappropriate if $R(t)$ recorded whether a subject received a mammogram or any other truly 'point-source' treatment at time t where $R(t)$ is a point-source exposure if $\{t; R(t) \neq 0\}$ is finite w.p. 1. Models for the effect of repeated point-source treatments, such as mammography, are discussed in Ref. 23, App. 2.

It then follows from Assumption 4 and Theorem (2.3) of Section 6 of Loomis and Sternberg³² that: (i) there exists a unique continuous solution $H(t) \equiv h(t, Y, \bar{L}(K + 1), \bar{R}(K + 1))$ to the differential equation $dH(t)/dt = D(H(t), t)$ satisfying $H(K + 1) = Y$; and (ii) the solution is a continuous function of (Y, t) on $R \times [0, K + 1]$ and $\partial h(t, Y, \bar{L}(K + 1), \bar{R}(K + 1)) / \partial(Y, t)$ exists and is bounded and continuous on $R \times [T_m, T_{m+1})$. Hence the Jacobian $\partial H(0) / \partial Y$ for the transformation from Y to $H(0)$ exists with probability 1. The strong smoothness conditions in Assumption 4 (ii) guarantee the existence of this Jacobian. Our main result is Theorem (8.2). In treatment arm $Z = d$:

Theorem 8.2: $H(t)$ and $Y_{(\bar{R}(t^-), 0)}$ have the same conditional distribution given $(\bar{L}(t), \bar{R}(t))$. In particular, $H \equiv H(0)$ and $Y_{(0)} \equiv Y(d)$ have the same marginal distribution.

Theorem (8.2) is proved in the Appendix for the special case in which there is local rank preservation, that is, $Q(Y_{\bar{R}(t^-), 0}, t, h) = Y_{(\bar{R}(t+h^-), 0)}$ and $Y_{(\bar{R}(t^-), 0)}$ is continuous in t with probability one and for the case where the number of jumps and the jump times are non-random. Although I am nearly certain that Theorem (8.2) holds with random jump times and no local rank preservation, my current proof attempt still suffers from unresolved technical problems.

We say the data follows a continuous time SNDM $D(y, t, \psi)$ if $D(y, t) = D(y, t, \psi_0)$ where ψ_0 is an unknown parameter and $D(y, t, \psi) \equiv D(y, t, \bar{L}(t), \bar{R}(t), \psi)$ is a known function satisfying (i) $D(y, t, 0) = 0$, (ii) $D(y, t, \psi) = 0$ if $R(t) = 0$, and (iii) Assumption 4(ii) holds for each fixed value of ψ . It then follows if the explainable non-random non-compliance assumption (10) holds, the Cox model (12) for C and the regression model $h_{g^*}[\bar{L}(C^-), \alpha]$ for $G^*(C)$, and the continuous time SNDM model $D(y, t, \psi)$ are correctly specified, then the estimators $\hat{\psi}_{ge}, \hat{\psi}_{ge, final}, \hat{\beta}_{ge}$ and $\hat{\beta}_{ge, final}$ will be consistent for ψ_0 and $E[Y(d)]$, respectively, even when the function $g(H(\psi), \bar{L}(t^-))$ to be added to the Cox model (12) is not linear in $H(\psi)$.

9. SENSITIVITY ANALYSIS FOR NON-IGNORABLE NON-COMPLIANCE

The methods in this section are in the spirit of but differ from methods developed by Rosenbaum in reference 40.

9.1. Joint estimation under non-ignorable non-compliance

Returning to the discrete time setting, our fundamental assumption (3) of explainable non-random non-compliance is equivalent to

$$\text{pr}[R_m = 0 | \bar{R}_{m-1} = \mathbf{0}, \bar{L}_m, Y(d)] = \text{pr}[R_m = 0 | \bar{R}_{m-1} = \mathbf{0}, \bar{L}_m]. \tag{14}$$

To allow for the possibility that the left-hand side of (14) actually does depend on $Y(d)$ (that is, we have non-ignorable non-compliance given \bar{L}_m), we specify a logistic model

$$\text{logit } \pi(m, \alpha) = \alpha_{1m} + \alpha'_2 h_m(\bar{L}_m) + \alpha'_3 q_m(Y(d), \bar{L}_m) \tag{15}$$

for the left-hand side of (14) with $\alpha = (\alpha'_1, \alpha'_2, \alpha'_3)'$, $\alpha'_1 = (\alpha_{10}, \dots, \alpha_{1K})$ where $h_m(\cdot)$ is a known vector function of \bar{L}_m and $q_m(\cdot, \cdot)$ is a known (possibly) vector valued function of $Y(d)$ and \bar{L}_m (for example, $q_m(Y(d), \bar{L}_m)$ could just be $Y(d)$). Let α^* be the true value of α , that is, $\text{pr}[R_m = 0 | \bar{L}_m, Y(d), \bar{R}_{m-1} = \mathbf{0}] \equiv \pi(m) = \pi(m, \alpha^*)$.

Remark: When we believe that compliance may be non-ignorable given our covariates \bar{L}_m , the model (15) will be hard to specify since, as discussed in the Remark in Section 3.4, $Y(d)$ is not observed for subjects with $C < K + 1$. Thus, at a minimum, I would recommend, as a sensitivity analysis, repeating the analysis described below for a number of different choices of the function $q_m(Y(d), \bar{L}_m)$, corresponding to different assumptions about the functional form of the effect of $Y(d)$ on compliance given \bar{L}_m . An alternative approach, which I favour as being more honest, is described in the following subsection.

Given model (15), the G-estimate $\hat{\psi}_{ge}$ of ψ_0 of a structural nested distribution model is calculated just as in Section 4.2 except now the extended logistic model is

$$\text{logit } \pi(m, \alpha, \theta) = \alpha_{1m} + \alpha'_2 h_m(\bar{L}_m) + \alpha'_3 q_m(H(\psi), \bar{L}_m) + \theta' g_m(H(\psi), \bar{L}_m) \tag{16}$$

where $g_m(\cdot, \cdot)$ is a (possibly vector-valued) function of the dimension of ψ chosen by the investigator.^{33,36} A 95 per cent confidence set for ψ_0 is the set of ψ for which a 5 per cent level likelihood-based test of hypothesis $\theta = 0$ does not reject.

Remark: We note that the above approach to sensitivity analysis will fail for SNMMs since, in contrast to an SNDM, the conditional distribution of $R_m | \bar{R}_{m-1} \equiv 0, \bar{L}_m, Y(d)$ need not equal the conditional distribution of $R_m | \bar{R}_{m-1} \equiv 0, \bar{L}_m, H(\psi_0)$. Thus, to conduct a sensitivity analysis for an SNMM model, we must: (i) replace $Y(d)$ by $H(\psi_0)$ in model (15) (since under an SNMM, $Y(d)$ and $H(\psi_0)$ are not fungible); (ii) choose $g_m(H(\psi), \bar{L}_m)$ in (16) linear in $H(\psi)$ to obtain a g-estimator $\hat{\psi}_{ge}$, and (iii) use the model in reference 12, p. 2396 to obtain a confidence interval for ψ_0 . Throughout the remainder of this subsection and the next, we restrict attention to SNDMs.

With non-ignorable missingness (that is, when (3) is false but model (15) is true), the true value of ψ_0 may not be identified, the asymptotic variance of $\hat{\psi}_{ge}$ may be infinite (and, for example, for an SNDM, the likelihood ratio score and Wald test of the hypothesis $\theta = 0$ may not reject for any value of ψ) and there may not be a unique estimate $\hat{\psi}_{ge}$. We now give a sufficient *theoretical* condition that guarantees the estimate $\hat{\psi}_{ge}$ of ψ_0 in a SNDM will indeed have infinite asymptotic variance. We then describe how this theoretical result might be used in practice.

Let $S_\psi(\psi, \alpha) = \partial \log \mathcal{L}(\psi, \alpha) / \partial \psi$ be the score for ψ for a single subject. Based on the likelihood

$$\begin{aligned} \mathcal{L}(\psi, \alpha) = & \{ \partial H(\psi) / \partial Y \} f \{ H(\psi) \} \prod_{m=0}^{\min(C, K)} f(L_m | \bar{L}_{m-1}, \bar{R}_{m-1} \equiv \mathbf{0}, H(\psi)) \\ & \times \prod_{m=0}^C f(R_m | \bar{R}_{m-1} \equiv \mathbf{0}, \bar{L}_m, H(\psi); \alpha). \end{aligned} \quad (17)$$

Note in practice, $S_\psi(\psi, \alpha)$ is a theoretical object since we have not specified models for $f\{H(\psi)\}$ or $f(L_m | \bar{L}_{m-1}, \bar{R}_{m-1} \equiv \mathbf{0}, H(\psi))$. Let $g_{opt,m}(H(\psi_0), \bar{L}_m)$ be the unknown function

$$E[S_\psi(\psi_0, \alpha^*) | \bar{L}_m, \bar{R}_{m-1} \equiv \mathbf{0}, H(\psi_0), R_m = 1] - E[S_\psi(\psi_0, \alpha^*) | \bar{L}_m, \bar{R}_{m-1} \equiv \mathbf{0}, H(\psi_0), R_m = 0].$$

Then $\hat{\psi}_{ge}$ that uses $g_{opt,m}[H(\psi), \bar{L}_m]$ is the most efficient G-estimator. In fact, it attains the semi-parametric variance bound^{37,38} for the semi-parametric model characterized by the SNDM $H(\psi)$ and the model (15), but which leaves the law of $H(\psi_0)$ (equivalently $Y(d)$), the law of L_m given $\bar{L}_{m-1}, \bar{R}_{m-1} \equiv \mathbf{0}, H(\psi_0)$ (or equivalently $Y(d)$), and the law of (\bar{L}_K, \bar{R}_K) given $C, \bar{L}_C, H(\psi_0)$ (equivalently, $Y(d)$) all completely unspecified. Hence, if the asymptotic variance of $\hat{\psi}_{ge}$ that uses $g_{opt,m}$ is not finite, then no G-estimator has a finite asymptotic variance. A necessary and sufficient condition for $\hat{\psi}_{ge}$ that uses $g_{opt,m}$ to have infinite asymptotic variance is that, when $\psi = \psi_0$, for each m and each subject, $g_{opt,m}[H(\psi), \bar{L}_m]$ is a linear combination of the regressors $(1, h_m(\bar{L}_m), q_m(H(\psi), \bar{L}_m))'$ in model (15). In particular, if $q_m(H(\psi), \bar{L}_m)$ equals $g_{opt,m}(H(\psi), \bar{L}_m)$, then all possible G-estimators $\hat{\psi}_{ge}$ will have infinite asymptotic variance.

We now consider how an analyst might use this theoretical result in practice. Suppose we obtain a confidence interval based on some initial chosen function $g_m(H(\psi), \bar{L}_m)$. If this interval is reasonably narrow, then we have carried out a successful G-analysis. However, if our 95 per cent confidence interval for ψ is too wide to be substantively useful, then either (i) our choice of the function g_m was quite inefficient or (ii) no choice of g_m , including the optimal choice $g_{opt,m}$, would give usefully narrow intervals. The best approach to try to discriminate between explanations (i) and (ii) is as follows. As when constructing the estimator $\hat{\beta}_{mar}$ in Section 4, specify parametric

models, depending on parameter $\gamma = (\gamma'_1, \gamma'_2)$ for the unparameterized densities in $\mathcal{L}(\psi, \alpha)$ and rewrite $\mathcal{L}(\psi, \alpha)$ as $\mathcal{L}(\psi, \alpha, \gamma)$ to reflect this dependence, find the maximum likelihood estimators $(\hat{\psi}, \hat{\gamma}, \hat{\alpha})$ based on maximizing the product over the n subjects of the $\mathcal{L}(\psi, \alpha, \gamma)$, and finally construct an estimate $\hat{g}_{\text{opt}, m}(H(\hat{\psi}), \bar{L}_m)$ of $g_{\text{opt}, m}$ based on the distribution implied by $(\hat{\psi}, \hat{\gamma}, \hat{\alpha})$. Now construct a new confidence interval based on G-estimation using the function $\hat{g}_{\text{opt}, m}(H(\hat{\psi}), \bar{L}_m)$ rather than the original $g_m(H(\psi), \bar{L}_m)$. The resulting estimator and 95 per cent confidence interval is said to be a locally efficient G-estimate and interval (at the parametric submodel indexed by γ). It is a valid confidence interval for ψ_0 even when the models parameterized by γ are misspecified. If this 95 per cent confidence interval for ψ_0 is reasonably narrow, we report this interval and conclude that option (i) above was true.

If the resulting confidence interval is still uselessly wide, we conclude it is likely that there is insufficient estimate information about ψ_0 in our semi-parametric model which leaves the marginal density of $H(\psi_0)$ (equivalently, of $Y(d)$) and the conditional law L_m given \bar{L}_{m-1} and $H(\psi_0)$ unspecified. In that case we might report an estimator which is the non-ignorable version of $\hat{\beta}_{\text{mar}}$ which will always be more efficient than the estimator $\hat{\beta}_{\text{ge}}$. Specifically, given $(\hat{\psi}, \hat{\gamma}, \hat{\alpha})$ calculated as just described, $\hat{\beta}_{\text{mar}} = \int y f(y; \hat{\gamma}_1) dy$. $\hat{\beta}_{\text{mar}}$ will be consistent for $E[Y(d)]$ provided the SNDM $H(\psi)$, our non-ignorable non-compliance model (15), and the parametric models $f\{Y(d); \gamma_1\}$ and $f\{L_m | \bar{L}_{m-1}, \bar{R}_{m-1} \equiv \mathbf{0}, Y(d); \gamma_2\}$ are correctly specified. As described in the Remark in Section 3.4, results based on $\hat{\beta}_{\text{mar}}$ should be viewed with caution because of the difficulty in correctly specifying the parametric model for L_m given $\bar{L}_{m-1}, \bar{R}_{m-1} \equiv \mathbf{0}, Y(d)$.

The non-ignorable compliance model (15) can also be used to investigate the dependence of the IPCW estimator $\hat{\beta}_w$ on the assumption (3) of explainable non-random non-compliance. However, the computational details are somewhat complex, and we therefore refer the reader to references 13, 27 and 28.

9.2. A more honest approach to sensitivity analysis

We showed in Section 9.1 that an essential requirement for g-estimation of ψ_0 under the non-ignorable non-compliance model (15) was that the function $g_m(\cdot, \cdot)$ in (16) was not a linear combination of the elements of $q_m(\cdot, \cdot)$ and $h_m(\cdot)$. However, since, in truth, we have no real idea of the functional form of the dependence of $\pi(m)$ on $Y(d)$ as encoded in $q_m(Y(d), \bar{L}_m)$, I believe an honest approach to sensitivity analysis is to take $q_m(\cdot, \cdot)$ equal to $g_m(\cdot, \cdot)$. As a consequence, ψ_0 will not be estimable by g-estimation based on $g_m(\cdot, \cdot)$ when the coefficient α_3 of $q_m(\cdot, \cdot)$ is unknown, a desirable state of affairs if one believes it is dishonest to identify ψ_0 by the arbitrary choice of a functional form $q_m(\cdot, \cdot)$. Furthermore, since $q_m(\cdot, \cdot)$ is a function of the often unobserved $Y(d)$, there is no honest way to choose the dimension of α_2 or the functional form $h_m(\bar{L}_m)$ in the non-ignorable model (15). However, when L_k has continuous components or is high-dimensional, due to the curse of dimensionality, we cannot replace $\alpha'_2 h_m(\bar{L}_m)$ by an arbitrary unknown function that we would estimate by, say, multivariate kernel smoothing. As a compromise, I would suggest we ensure that $\alpha'_2 h_m(\bar{L}_m)$ is richly parametrized. To obtain a sensitivity analysis when $q_m = g_m$, I suggest the following algorithm:

- (i) select a richly parameterized model $\alpha'_2 h_m(\bar{L}_m)$ and what one believes to be a reasonably 'efficient' function $g_m(\cdot, \cdot)$ in (16) and set $q_m(\cdot, \cdot) = g_m(\cdot, \cdot)$;
- (ii) regard α_3 as a sensitivity parameter that we will vary (but not estimate) in the sensitivity analysis;

- (iii) estimate ψ_0 by g-estimation of model (16) with α_3 held fixed, thereby obtaining estimates $\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\psi}_{ge}$; and
- (iv) plot a graph of the g-estimate $\tilde{\psi}_{ge}$ and confidence interval for ψ_0 as a function of the sensitivity parameter α_3 .

The g-estimate of ψ_0 will be consistent and asymptotically normal for each fixed value of the selection parameter α_3 if model (15) is correctly specified. A drawback of the above algorithm is that the interpretation of the degree of selection bias as measured by α_3 will depend on the choice of the functions $q_m(\cdot, \cdot) = g_m(\cdot, \cdot)$. To make α_3 easily interpretable, one might therefore wish to choose $q_m(Y(d), \bar{L}_m) = Y(d)$ while to ensure efficiency of estimation (even with α_3 fixed), one may want $g_m(H(\psi_0), \bar{L}_m)$ to depend on \bar{L}_m . In that case, one can still use the above algorithm without enforcing $g_m = q_m$ but still treating α_3 as a sensitivity parameter that will be varied but not estimated. For each choice of α_3 , in order to obtain the narrowest possible intervals for ψ_0 , we would use the estimate $\hat{g}_{opt,m}(\cdot, \cdot)$ of $g_{opt,m}(\cdot, \cdot)$ described in Section 9.1. However, any attempt to do so encounters some subtle difficulties explored further in Appendix II.

10. CENSORING BY LOSS TO FOLLOW-UP

10.1. Explainable censoring

Heretofore, we have assumed that data on the outcome Y is obtained on all study subjects. In practice, some subjects will inevitably be lost to follow-up (censored) prior to the time $K + 1$ at which Y is recorded. Let us define the censoring time Q to be k if a subject was lost to follow-up after time k but before time $k + 1$. If a subject is not lost to follow-up, we set $Q = K + 1$. Thus Y is observed for subjects only for whom $Q = K + 1$. Suppose we can assume that censoring by Q is explainable by \bar{R}_k and \bar{L}_k , that is, in arm $Z = d$ the probability of being censored at k given \bar{L}_k and \bar{R}_k among those uncensored prior to k does not depend on the possibly unobserved value of Y . That is,

$$\text{pr}[Q = k | Q \geq k - 1, \bar{L}_k, \bar{R}_k, Y] = \text{pr}[Q = k | Q \geq k - 1, \bar{L}_k, \bar{R}_k]. \quad (18)$$

Now let the logistic model

$$\text{logit } \lambda(m, \phi) = \phi_{1m} + \phi'_2 h_m^*(\bar{L}_m, \bar{R}_m), \quad m = 0, \dots, K \quad (19)$$

be a model for the probability of *not* being censored at time m given \bar{L}_m, \bar{R}_m among those uncensored up to time m in arm $Z = d$, where $h_m^*(\bar{L}_m, \bar{R}_m)$ is a known vector-valued function of (\bar{L}_m, \bar{R}_m) , $\phi = (\phi'_1, \phi'_2)$, and $\phi'_1 = (\phi_{10}, \dots, \phi_{1K})$. Then the maximum likelihood estimator $\hat{\phi}$ maximizes the logistic likelihood $\prod_{i=1}^{n_d} \mathcal{L}_i^Q(\phi)$ where $\mathcal{L}^Q(\phi) = \prod_{m=0}^{Q-1} \lambda(m, \phi) [1 - \lambda(Q, \phi)]^{I(Q=K+1)}$ and n_d is the number of subjects in arm $Z = d$.

We then modify our previous estimation procedures to accommodate censoring by Q as follows. The IPCW estimator becomes

$$\hat{\beta}_w = \left[\sum_{i=1}^{n_d} I(Q = K + 1) \delta_i Y_i / \{ \bar{\pi}_i(K, \hat{\alpha}) \bar{\lambda}_i(K, \hat{\phi}) \} \right] / \left[\sum_{i=1}^{n_d} I(Q = K + 1) \delta_i / \{ \bar{\pi}_i(K, \hat{\alpha}) \bar{\lambda}_i(K, \hat{\phi}) \} \right]$$

where $\bar{\lambda}(m, \phi) = \prod_{k=0}^m \lambda(k, \phi)$. That is, subjects uncensored by loss to follow-up who also have $\bar{R}_K = \mathbf{0}$ contribute to a weighted sum of the Y 's with weight proportional to $\{ \bar{\pi}(K, \hat{\alpha}) \bar{\lambda}(K, \hat{\phi}) \}^{-1}$.

Under a slight strengthening of (18), $\bar{\lambda}_i(K, \hat{\phi})$ is an estimate of the conditional probability of subject i not having been lost to follow-up through time $K + 1$.¹³

$\hat{\psi}_{ge}$ remains as defined as in Section 4.3 except the weighted ‘likelihood’ $\mathcal{L}_{mis}^{*,Q}(\alpha, \theta) = \mathcal{L}_{mis}^*(\alpha, \theta)^{I(Q=K+1)/\bar{\lambda}(K, \hat{\phi})}$ replaces $\mathcal{L}_{mis}^*(\alpha, \theta)$. That is, only subjects uncensored by end to follow-up (that is, those for whom Y is observed) contribute to the ‘likelihood’, and their contribution to the logarithm of the likelihood is weighted by $1/\bar{\lambda}(K, \hat{\phi})$. Then

$$\hat{\beta}_{ge} \equiv \left[\sum_{i=1}^{n_d} H_i(\hat{\psi}_{ge}) I(Q = K + 1) / \bar{\lambda}_i(K, \hat{\phi}) \right] / \left[\sum_{i=1}^{n_d} I(Q = K + 1) / \bar{\lambda}_i(K, \hat{\phi}) \right]$$

becomes a weighted sum of the $H_i(\hat{\psi}_{ge})$ among subjects for whom Y is observed.

$\hat{\beta}_g, \hat{\beta}_{mar}$ and $\hat{\beta}_{mar}^*$ remain as defined previously except the likelihoods $\mathcal{L}_g(\theta_1, \theta_2), \mathcal{L}_{mar}(\gamma_1, \gamma_2)$, and $\mathcal{L}_{mar}^*(\gamma_1, \gamma_2, \psi)$ are replaced by their weighted likelihoods. For example, $\mathcal{L}_g(\theta_1, \theta_2)$ is replaced by $\mathcal{L}_g(\theta_1, \theta_2)^{I(Q=K+1)/\bar{\lambda}(K, \hat{\phi})}$.

The iterated conditional expectation estimator $\hat{\beta}_1$ of Section 4.2 is modified in that we now first perform a weighted regression with weights $1/\bar{\lambda}(K, \hat{\phi})$ of Y on (\bar{L}_K, \bar{R}_K) with regression function $h_K(\bar{L}_K, \bar{R}_K, \psi)$ among subjects who were not lost to follow-up ($Q = K + 1$). Then, for $K - 1, \dots, 0$, we let $\hat{\psi}_k$ be the weighted least squares estimate from the regression of $h_{k+1}(\bar{L}_{k+1}, \bar{R}_{k+1} = \mathbf{0}, \hat{\psi}_{k+1})$ on \bar{L}_k, \bar{R}_k with regression function $h_k(\bar{L}_k, \bar{R}_k, \psi_k)$ with weights $1/\bar{\lambda}(k, \hat{\phi})$ among subjects who were uncensored through $k + 1$, that is, $Q \geq k + 1$.

Confidence intervals for $\hat{\beta}_{mar}, \hat{\beta}_{mar}^*$ and $\hat{\psi}_{ge}$ can no longer be based on likelihood-based score, likelihood ratio, or Wald tests. Appropriate confidence procedures are described in references 12 and 13. We say that censoring by loss to follow-up is at random if the value of ϕ_2 in model (19) is zero. Only $\hat{\beta}_w$ is guaranteed to be consistent for $E[Y(d)]$ when censoring and non-compliance are both at random. Only $\hat{\beta}_w$ and $\hat{\beta}_{ge}$ are guaranteed to be consistent for $E[Y(d)]$ when censoring and non-compliance are at random and the null hypothesis (2) is true.

Remark: The above modifications of the estimators $\hat{\beta}_g, \hat{\beta}_1, \hat{\beta}_{mar}$ and $\hat{\beta}_{mar}^*$ in the presence of explainable censoring are not necessarily the only or even the most efficient possible modifications. For example, a second modification to $\hat{\beta}_g$ of Section 4.1 to accommodate censoring by loss to follow-up is to modify $\mathcal{L}_g(\theta_1, \theta_2)$ by using densities that condition on ‘not being censored’. That is, we modify $\mathcal{L}_g(\theta_1, \theta_2)$ to

$$\prod_{m=0}^K f[L_m | \bar{L}_{m-1}, \bar{R}_{m-1}, Q \geq m; \theta_1] f[Y | \bar{L}_K, \bar{R}_K, Q = K + 1; \theta_2].$$

This second modification will, under correct specification of the above parametric models, give a consistent estimator of $E[Y(d)]$ that is not only more efficient than the estimator obtained under the weighted likelihood modification, but also does not require us to specify a model (19) for censoring.

10.2. Non-ignorable censoring

To examine the sensitivity of these estimates of $E[Y(d)]$ to the assumption (18) of explainable censoring or to estimate $E[Y(d)]$ under a non-ignorable censoring model, we add the term $\phi'_3 q_m^*(\bar{L}_{m-1}, \bar{R}_{m-1}, Y)$ to model (19) so that the probability of being censored at time m now may depend on the possibly unobserved value of Y .

The modified estimators of Section 10.1 are still appropriate with $\hat{\phi}$ now being a consistent estimator of $\phi = (\phi'_1, \phi'_2, \phi'_3)$. However, estimation of ϕ must be carried out using the methods

described in references 13, 27, 41 and 28. As discussed previously, these methods are outside the scope of this paper, and we refer the reader to the above references.

11. CONCLUSION

In summary, I have presented a large number of methods for analysing equivalence trials with non-compliance. Although I have argued most forcefully for two semi-parametric methods – inverse probability of censoring weighted estimators, weighted GEE estimator of marginal structural models, and G-estimation of structural nested models (or continuous time structural nested models) – my aim has not been to preclude use of the other approaches. Since the analysis of trials with non-compliance is sensitive to assumptions, I believe the more analytic methods used, the better, as long as one summarizes and comments both on the consistency of the results and on the strengths and weaknesses of the assumptions underlying each method.

APPENDIX I

Proof of Theorem 8.1

Let $M(u, t) = \lim_{\Delta t \downarrow 0} \{Z(u + \Delta t, t) - Z(u, t)\} / \Delta t$ where $Z(u, t) \equiv E[Y_{(\bar{R}(u^-, 0))} | \bar{L}(t), \bar{R}(t)]$. Note $M(u, t) = E[D(u) | \bar{L}(t), \bar{R}(t)]$ and thus, by Assumption 3, is continuous except at T_m , $m = 1, \dots, K^*$. Hence

$$\begin{aligned} E[H(t) | \bar{L}(t), \bar{R}(t)] &= E[Y | \bar{L}(t), \bar{R}(t)] - E\left[\int_t^{K+1} D(u) du | \bar{L}(t), \bar{R}(t)\right] \\ &= E[Y | \bar{L}(t), \bar{R}(t)] - \left\{\int_t^{K+1} M(u, t) du\right\} \\ &= E[Y | \bar{L}(t), \bar{R}(t)] - \{E[Y | \bar{L}(t), \bar{R}(t)] + \sum_{\{m: T_m > t\}} \{Z(T_m^-, t) - Z(T_m, t)\} - Z(t, t)\} \\ &= Z(t, t) = E[Y_{(\bar{R}(t^-, 0))} | \bar{L}(t), \bar{R}(t)] \end{aligned}$$

where the last equality is definitional, the second to last is by the fact that, by Assumption 2, $Z(u, t)$ is continuous in u , so $Z(T_m^-, t) = Z(T_m, t)$, the third to last is by the facts that $M(u, t) = \partial Z(u, t) / \partial u$ for $u \in [T_m, T_{m+1})$ and $Z(K + 1, t) = E[Y | \bar{L}(t), \bar{R}(t)]$ by Consistency Assumption 1b.

Proof of Theorem 8.2 under Local Rank Preservation

By Theorem (2.3) of Chapter 6 of Loomis and Sternberg, there is a unique continuous solution to $dH(t)/dt = D(H(t), t)$, $t \in [T_m, T_{m+1})$ satisfying $H(K + 1) = Y$. We now show that under local rank preservation $Y_{(\bar{R}(t^-, 0))}$ satisfies these conditions as a function of t . First by Assumption (4) and local rank preservation (i) $d(Y_{(\bar{R}(t^-, 0))})/dt$ exists and equals $D(Y_{(\bar{R}(t^-, 0))}, t)$ on $[T_m, T_{m+1})$ with $Y_{(\bar{R}(t^-, 0))}$ continuous in t , and (ii) $(Y_{(\bar{R}(t^-, 0))})$ evaluated at $t = K + 1$ is Y by Consistency Assumption 1b.

Proof of Theorem 8.2 with Non-random Jump Times

We shall need some additional definitions and a lemma.

Let $V(y, t)$ be the unique random function from $R^1 \times [0, K + 1] \rightarrow R$ satisfying for $t \in [T_m, T_{m+1}), m = 0, \dots, K^*$,

$$\text{pr}[Y_{(\bar{R}(T_m), 0)} > y | \bar{L}(T_m), \bar{R}(T_m)] = \text{pr}[Y_{(\bar{R}(t^-), 0)} > V(y, t) | \bar{L}(T_m), \bar{R}(T_m)].$$

Note $V(y, T_m) = y$ but $V(y, T_m^-) = y$ where $V(y, T_m^-) = \lim_{u \uparrow T_m} V(y, u)$.

Since $V(y, t)$ is increasing in y for fixed t , the inverse function $V^{-1}(u, t)$ exists where $V^{-1}(u, t) \equiv y$ if $V(y, t) = u$. Let $D^*(y, t) = \lim_{\Delta t \downarrow 0} \{V(y, t + \Delta t) - V(y, t)\} / \Delta t$. Then we have the following lemma.

Lemma: $D(y, t) = D^*(V^{-1}(y, t), t)$.

Proof: From the definition of $Q(y, t, h)$, we have

$$V^{-1}(Q(y, t, h), t + h) = V^{-1}(y, t) \tag{20}$$

since for T_m such that $T_m \leq t < T_{m+1}$, the left hand side of (13) equals

$$\text{pr}[Y_{(\bar{R}(T_m), 0)} > V^{-1}(Q(y, t, h), y + h) | \bar{L}(T_m), \bar{R}(T_m)]$$

and the right hand side of (13) equals

$$\text{pr}[Y_{(\bar{R}(T_m), 0)} > V^{-1}(y, t) | \bar{L}(T_m), \bar{R}(T_m)]$$

where we have used the fact that when the T_m are non-random the event $\bar{L}(T_m), \bar{R}(T_m)$ is the same event as $\bar{L}(t), \bar{R}(t)$. Now differentiating both sides of (20) with respect to h , we obtain for $t \in [T_m, T_{m+1}), V_1^{-1}(y, t) \partial Q(y, t, h) / \partial h + V_2^{-1}(y, t) = 0$ where $V_k^{-1}(\cdot, \cdot)$ refers to the partial derivative of $V^{-1}(\cdot, \cdot)$ with respect to its k^{th} argument. Hence

$$D(y, t) \equiv \partial Q(y, t, h) / \partial h = -V_2^{-1}(y, t) / V_1^{-1}(y, t). \tag{21}$$

However, by differentiating $V^{-1}(V(x, t), t) = x$ with respect to t , we obtain $V_1^{-1}(V(x, t), t) \partial V(x, t) / \partial t + V_2^{-1}(V(x, t), t) = 0$. Thus

$$D^*(x, t) \equiv \partial V(x, t) / \partial t = -V_2^{-1}(V(x, t), t) / V_1^{-1}(V(x, t), t).$$

The result now follows upon substituting $V^{-1}(y, t)$ for x in the last display and comparing with (21).

Corollary: $H(t) = V[H(T_m), t]$ for $T_m \leq t < T_{m+1}$. Further, $H(T_m) = V^{-1}[H(T_{m+1}), T_m^-]$.

Proof: Since, (i) by the previous lemma, for $t \neq T_m \partial V(y, t) / \partial t = D(V(y, t), t)$, (ii) by continuity of $H(t), Y = H(T_{K^*+1})$ and, (iii) by assumption 2, $V\{V^{-1}(Y, T_{K^*+1}), T_{K^*+1}\} = Y$, we have, by the uniqueness of solutions to differential equations through a common point (Theorem 2.3 of Chapters 6 of Loomis and Sternberg), that, for $t \in [T_{K^*}, T_{K^*+1}), H(t) = V\{V^{-1}(Y, T_{K^*+1}), t\}$. In particular, by substituting T_{K^*} for $t, H(T_{K^*}) = V^{-1}(Y, T_{K^*+1})$. Similarly, since $H(T_{K^*}) = H(T_{K^*})$ by continuity of $H(t)$, and $V\{V^{-1}[H(T_{K^*}), T_{K^*}], T_{K^*}\} = H(T_{K^*})$, we again have by the uniqueness of solutions to differential equations that for $t \in [T_{K^*-1}, T_{K^*}), H(t) = V[V^{-1}\{H(T_{K^*}), T_{K^*}\}, t]$ and $H(T_{K^*-1}) = V^{-1}[H(T_{K^*}), T_{K^*}]$. Continuing in this fashion proves the corollary.

Proof of Theorem 8.2: By the corollary and definition of $V(y, t)$, it is enough to prove the theorem for $t = T_m, m = K^* + 1, \dots, 0$ by induction. Now $H(T_{K^*+1})$ and $Y_{(\bar{R}(T_{K^*+1}^-), 0)}$ are equal to Y and thus to one another by definition. Hence it remains to be shown that if we assume that the theorem is true for T_m , then it is true for T_{m-1} . If the theorem is true for T_m , then, by T_m non-random,

$$\text{pr}[H(T_m) > x | \bar{L}(T_{m-1}), \bar{R}(T_{m-1})] = \text{pr}[Y_{\bar{R}(T_m^-, 0)} > x | \bar{L}(T_{m-1}), \bar{R}(T_{m-1})].$$

Now the left hand side of the last display equals

$$\lim_{u \uparrow T_m} \text{pr}[H(u) > x | \bar{L}(T_{m-1}), \bar{R}(T_{m-1})]$$

by dominated convergence and continuity of $H(t)$; and, therefore, also equals

$$\lim_{u \uparrow T_m} \text{pr}[H(T_{m-1}) > V^{-1}(x, u) | \bar{L}(T_{m-1}), \bar{R}(T_{m-1})]$$

by the previous corollary. But

$$\text{pr}[Y_{(\bar{R}(T_m^-), 0)} > x | \bar{L}(T_{m-1}), \bar{R}(T_{m-1})]$$

equals

$$\lim_{u \uparrow T_m} \text{pr}[Y_{(\bar{R}(u^-), 0)} > x | \bar{L}(T_{m-1}), \bar{R}(T_{m-1})]$$

by dominated convergence and Assumption 2, and therefore, also equals

$$\lim_{u \uparrow T_m} \text{pr}[Y_{(\bar{R}(T_{m-1}^-), 0)} > V^{-1}(x, u) | \bar{L}(T_{m-1}), \bar{R}(T_{m-1})]$$

by definition of $V(y, t)$. However, as x varies, $V^{-1}(x, T_m^-)$ varies over all of R^1 proving the theorem.

APPENDIX II

In this Appendix, we consider whether one can carry out the sensitivity analysis algorithm of Section 9.2 with the function g_m equal to the optimal function $g_{\text{opt}, m}$, under a generalization of the non-compliance model (15) that allows the last term in model (15) to be non-linear in α_3 . That is, we replace (15) by

$$\text{logit } \pi(m, \alpha) = \alpha_{1m} + \alpha_2 h_m(\bar{L}_m) + q\{H(\psi), \bar{L}_m, \alpha_3\}. \quad (22)$$

Mathematically, the question we are then asking is does there exist a function $q(\cdot, \cdot, \cdot)$ such that for all values of α_3 , the information bound for ψ_0 is (i) zero in the semi-parametric model characterized by the SNDM $H(\psi)$ and the non-compliance model (22) with α_3 unknown, but (ii) the information for ψ_0 is non-zero for α_3 known. Since this question is largely of theoretical rather than practical interest, we only consider the following special case in which $K = 0, H(\psi) = Y - \psi R_0$. To simplify notation, set $R \equiv R_0, L \equiv L_0, H \equiv H(\psi_0), s(H, L) = \partial \ln f(H | L) / \partial H$ and $\pi = \pi(H, L, \alpha) = \text{expit}[\alpha_1 + \alpha_2 h(L) + q(H, L, \alpha_3)]$ with $\text{expit}(x) = e^x / (1 + e^x)$. Then the likelihood function is $\mathcal{L}(\psi, \alpha) = f(L) f\{H(\psi) | L\} f(R | L, H(\psi); \alpha)$. Thus the score for ψ evaluated at

the truth is $S_\psi = -s(H, L)R - R[1 - \pi]\{\partial q(H, L, \alpha_3)/\partial H\}$ since

$$(i) \quad \partial H(\psi)/\partial \psi = -R \text{ and}$$

$$(ii) \quad \partial/\partial \psi \{ \ln[\exp\{q(H(\psi), L, \alpha_3)R\}/\{1 + \exp[q(H(\psi), L, \alpha_3)]\}] \} \\ = (R - \pi) [\partial q(H, L, \alpha_3)/\partial H]/[\partial H(\psi)/\partial \psi].$$

The infinite dimensional parameter in our semi-parametric model is the law of $H|L$. The vector $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ and ψ are the finite dimensional parameters. The nuisance tangent space $\Lambda = \{a(H, L); E[a(H, L)|L] = 0\}$ are functions of H and L with mean zero given L . Then the efficient score $S_{\psi, \text{eff}}$ for ψ is the residual of the projection of S_ψ on Λ

$$S_{\psi, \text{eff}} = S_\psi - \{E(S_\psi|H, L) - E[S_\psi|L]\} = (R - \pi)[-s(H, L) - \{\partial q(H, L, \alpha_3)/\partial H\}(1 - \pi)].$$

Similarly, we calculate the score and efficient score for α_3 to be

$$S_{\alpha_3} = (R - \pi) \partial q(H, L, \alpha_3)/\partial \alpha_3 = S_{\alpha_3, \text{eff}}.$$

It follows that we will have no information for ψ when α_3 is unknown if and only if $S_{\psi, \text{eff}} = cS_{\alpha_3}$ for some constant c . This is equivalent to saying that $q(H, L, \alpha_3)$ solves the partial differential equation $-s(H, L) - [\partial q(H, L, \alpha_3)/\partial H][1 - \pi(H, L, \alpha)] = c \partial q(H, L, \alpha_3)/\partial \alpha_3$ for some constant c .

ACKNOWLEDGEMENTS

Support for this research was provided in part by grants 2 P30 ES00002, RO1-A132475, RO1-ESO3405, K04-ES00180, GM-48704 and GM-29745 from the National Institutes of Health.

REFERENCES

1. Robins, J. M. '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 (1986).
2. Robins, J. M. '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 (1987).
3. Robins, J. M. 'The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies', in Sechrest, L., Freeman, H., Mulley, A. (eds), *Health Service Research Methodology: A Focus on AIDS*, NCHSR, U.S. Public Health Service, 1989, pp. 113–159.
4. Robins, J. M. and Tsiatis, A. 'Correcting for non-compliance in randomized trials using rank-preserving structural failure time models', *Communications in Statistics*, **20**, 2609–2631 (1991).
5. Robins, J. M. and Rotnitzky, A. 'Recovery of information and adjustment for dependent censoring using surrogate markers', in Jewell, N., Dietz, K., Farewell, V. (eds.), *AIDS Epidemiology – Methodological Issues*, Birkhäuser, Boston, MA, 1992, pp. 297–331.
6. Robins, J. M. 'Estimation of the time-dependent accelerated failure time model in the presence of confounding factors', *Biometrika*, **79**, 321–334 (1992).
7. Robins, J. M. 'Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers', *Proceedings of the Biopharmaceutical Section, American Statistical Association*, 24–33 (1993).
8. Robins, J. M. 'Analytic methods for estimating HIV treatment and cofactor effects', in Ostrow, D. G. and Kessler, R. (eds), *Methodological Issues of AIDS Mental Health Research*, Plenum Publishing, New York, 1993, pp. 213–290.

9. Mark, S. D. and Robins, J. M. 'Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model', *Statistics in Medicine*, **12**, 1605–1628 (1993).
10. Mark, S. D. and Robins, J. M. 'A method for the analysis of randomized trials with compliance information: An application to the multiple risk factor intervention trial', *Controlled Clinical Trials*, **14**, 79–97 (1993).
11. Robins, J. M. and Greenland, S. 'Adjusting for differential rates of PCP prophylaxis in high- versus low-dose AZT treatment arms in an AIDS randomized trial', *Journal of the American Statistical Association*, **89**, 737–749 (1994).
12. Robins, J. M. 'Correcting for non-compliance in randomized trials using structural nested mean models', *Communications in Statistics*, **23**, 2379–2412 (1994).
13. Robins, J. M., Rotnitzky, A. and Zhao, L.-P. 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of the American Statistical Association*, **90**, 106–121 (1995).
14. Robins, J. M. and Rotnitzky, A. 'Semiparametric efficiency in multivariate regression models with missing data', *Journal of the American Statistical Association*, **90**, 122–129 (1995).
15. Rotnitzky, A. and Robins, J. M. 'Semiparametric regression estimation in the presence of dependent censoring', *Biometrika*, **82**, 805–820 (1995).
16. Angrist, J. D., Imbens, G. W. and Rubin, D. B. 'Identification of causal effects using instrumental variables', *Journal of the American Statistical Association*, **91**, 444–455 (1995).
17. Heckman, J. J. and Robb, R. 'Alternative methods for evaluating the impact of interventions', in Heckman, J. J. and Singer, B. (eds), *Longitudinal Analysis of Labor Market Data*, Cambridge University Press, London, 1985, pp. 156–246.
18. Heyting, A., Tolboom, J. T. B. M. and Essers, J. G. A. 'Statistical handling of drop-outs in longitudinal clinical trials', *Statistics in Medicine*, **11**, 2043–2062 (1992).
19. Permutt, T. and Hebel, J. R. 'Simultaneous equation estimation in a clinical trial of the effect of smoking on birth weight', *Biometrics*, **45**, 619–622 (1989).
20. Robins, J. M., Blevins, D., Ritter, G. and Wulfsohn, M. 'G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients', *Epidemiology*, **3**, 319–336 (1992).
21. Balke, A. and Pearl, J. 'Non-parametric bounds on causal effects from partial compliance data', Technical Report R-199-J, University of California, Los Angeles, Computer Science Department, 1994.
22. Rosenbaum, P. R. and Rubin, D. B. 'The central role of the propensity score in observational studies for causal effects', *Biometrika*, **70**, 41–55 (1983).
23. Manski, C. F. 'Learning about social programs from experiments with random assignment of treatments', *SSRI* 9505, 1994.
24. Rubin, D. B. 'Bayesian inference for causal effects: the role of randomization', *Annals of Statistics*, **6**, 34–58 (1978).
25. Robins, J. M. and Greenland, S. 'Comment: The estimation of global average treatment effects using instrumental variables', *Journal of the American Statistical Association*, **91**, 456–458 (1996).
26. Robins, J. M. 'Discussion of "Causal diagrams for empirical research" by J. Pearl', *Biometrika*, **82**, 695–698 (1995).
27. Rotnitzky, A. and Robins, J. M. 'Analysis of semiparametric regression models with non-ignorable non-response', *Statistics in Medicine*, **16**, 81–102 (1997).
28. Rotnitzky, A. and Robins, J. M. 'Estimation of semiparametric regression models with non-ignorable missing data', (Submitted, 1995).
29. Breslow, N. E., Lubin, J. H., Marek, P. and Langholtz, B. 'Multiplicative models and cohort analysis', *Journal of the American Statistical Association*, **78**, 1–12 (1983).
30. Robins, J. M., Mark, S. D. and Newey, W. K. 'Estimating exposure effects by modelling the expectation of exposure conditional on confounders', *Biometrics*, **48**, 479–495 (1992).
31. Arjas, E. 'Survival models in martingale dynamics (with discussion)', *Scandinavian Journal of Statistics*, **15**, 177–225 (1989).
32. Loomis, B. and Sternberg, S. *Advanced Calculus*, Addison Wesley, 1968.
33. Robins, J. M. and Rotnitzky, A. 'Estimation of structural models with non-ignorable treatment assignment', (Submitted, 1996).

34. Robins, J. M., Greenland, S. and Hu, F.-C. 'Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome', submitted *Journal of the American Statistical Association, Applications and Case Studies* (1997).
35. Sommer, A. and Zeger, S. 'On estimating efficacy from clinical trials', *Statistics in Medicine*, **10**, 45–52 (1991).
36. Robins, J. M. 'Causal inference from complex longitudinal data', in Berkane, M. (ed.), *Latent Variable Modeling and Applications to Causality, Lecture Notes in Statistics* (120), Springer Verlag, New York, 1997, pp. 69–117.
37. Newey, W. K. 'Semiparametric efficiency bounds', *Journal of Applied Econometrics*, **5**, 99–135 (1990).
38. Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. *Efficient and Adaptive Inference in Semiparametric Models*, Johns Hopkins University Press, Baltimore, MD, 1993.
39. Goetghebeur, E. T. and Shapiro, S. H. 'Analyzing non-compliance in clinical trials: ethical imperative or mission impossible?', *Statistics in Medicine*, **15**, 2813–2826 (1996).
40. Rosenbaum, P. *Observational Studies*, Springer-Verlag, New York, 1995.
41. Robins, J. M. 'Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models', in: Cooper, G. and Glymour, C. (eds), *Causation and Computation*, MIT/AAAI Press (1998, to appear).
42. Robins, J. M. 'Marginal structural models', in: Halloran, E. (ed.), *Statistical Models in Epidemiology*. Springer-Verlag (1997, to appear).
43. Robins, J. M., Scharfstein, D. and Rotnitzky, A. 'Sensitivity analysis for selection bias in missing data and causal inference problems', in: Halloran, E. (ed.), *Statistical Models in Epidemiology*. Springer-Verlag (1997, to appear).