HHS Public Access

Author manuscript

Biometrika. Author manuscript; available in PMC 2016 July 21.

Published in final edited form as:

Biometrika. 2007 December ; 94(4): 841–860. doi:10.1093/biomet/asm070.

Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse

Stijn Vansteelandt,

Department of Applied Mathematics and Computer Sciences, Ghent University, 9000 Ghent, Belgium

Andrea Rotnitzky, and

Department of Economics, Di Tella University, Buenos Aires, Argentina

James Robins

Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.

Stijn Vansteelandt: stijn.vansteelandt@ugent.be; Andrea Rotnitzky: arotnitzky@utdt.edu; James Robins: robins@hsph.harvard.edu

Summary

We propose a new class of models for making inference about the mean of a vector of repeated outcomes when the outcome vector is incompletely observed in some study units and missingness is nonmonotone. Each model in our class is indexed by a set of unidentified selection bias functions which quantify the residual association of the outcome at each occasion t and the probability that this outcome is missing after adjusting for variables observed prior to time t and for the past nonresponse pattern. In particular, selection bias functions equal to zero encode the investigator's a priori belief that nonresponse of the next outcome does not depend on that outcome after adjusting for the observed past. We call this assumption sequential explainability. Since each model in our class is nonparametric, it fits the data perfectly well. As such, our models are ideal for conducting sensitivity analyses aimed at evaluating the impact that different degrees of departure from sequential explainability have on inference about the marginal means of interest. Although the marginal means are identified under each of our models, their estimation is not feasible in practice because it requires the auxiliary estimation of conditional expectations and probabilities given high-dimensional variables. We henceforth discuss estimation of the marginal means under each model in our class assuming, additionally, that at each occasion either one of following two models holds: a parametric model for the conditional probability of nonresponse given current outcomes and past recorded data, or a parametric model for the conditional mean of the outcome on the nonrespondents given the past recorded data. We call the resulting procedure 2^T-multiply robust as it protects at each of the T time points against misspecification of one of these two working models, although not against simultaneous misspecification of both. We extend our proposed class of models and estimators to incorporate data configurations which include baseline covariates and a parametric model for the conditional mean of the vector of repeated outcomes given the baseline covariates.

Some key words

Double robustness; Generalized estimating equation; Intermittent missingness; Longitudinal study; Missing at random; Semiparametric inference

1 Introduction

Consider a follow-up study whose design prescribes measurements of an outcome of interest to be taken on n independent subjects at fixed time-points. The goal of the study is to make inference about the mean outcome vector possibly as a function of baseline covariates and time. The intended vector of outcomes is often not completely recorded because some subjects miss some study cycles. When, as usual, the mechanism leading to these outcomes being missing is unknown to the investigator, the expected outcome at each time-point is not identified from the observed data. Inference must then rely on unverifiable assumptions about the missing data.

The problem of missing data in follow-up studies has received much attention in the statistical literature, but most emphasis has been given to settings where the missing pattern is monotone, in which no subject returns to subsequent study cycles after missing previous cycles. Robins et al. (1999) described a model for monotone missing data patterns which requires the a priori specification of a selection bias parameter that encodes the residual association between the outcome vector and missingness at each occasion after adjusting for past recorded data. They showed that, regardless of the value of the selection bias parameter, the model is nonparametric (just) identified as it imposes no restriction on the observed data distribution and yet identifies the mean of the repeated outcomes. Since all values of the selection bias parameter determine the same model for the observed data distribution, the selection bias parameter is not identified. Robins et al. (1999) therefore recommended conducting inference about the mean of the outcome vector by repeating the estimation under different plausible values for the selection bias parameter as a form of sensitivity analysis. The goal of this paper is to extend the work of Robins et al. (1999) to the case in which the outcome vector is incompletely observed in some study units and missingness is nonmonotone.

Our interest in nonparametric identified models is motivated by the fact that other models fail to distinguish (i) the nonidentifiable, i.e. untestable, restrictions on the missing data process necessary to identify the full-data parameter of interest from (ii) additional identifiable restrictions that serve to increase the efficiency of estimation. The distinction between (i) and (ii) is not only conceptually important but can also be practically important. For example, when one has available a nonparametric identified model, one can first fit the model to the data. If the resulting uncertainty concerning the functionals of interest, such as the marginal means of the outcomes, is too large to be of substantive use, as measured for instance by the volume of a 95% confidence region, then, as in any inferential problem, to reduce uncertainty one can choose between fitting a nested submodel and refitting the same model after collecting data on additional subjects, where possible, to increase the sample size. Clearly, when logistically and financially feasible, the second option is to be preferred.

Thus, fitting a model that is not nonparametrically identified is tantamount to supplementing additional modelling restrictions for the unavailable additional data. Furthermore, follow-up studies routinely collect high-dimensional data and models that are not nonparametrically identified require assumptions to be imposed on the mechanism generating these high-dimensional data. However, specification of realistic models is difficult, if not impossible. Nonparametric identified models meet the challenge posed by high-dimensional data because they only make assumptions about the missing data mechanism, thereby reducing the possibility of model misspecification.

Several available methods for the analysis of nonmonotone missing data assume that the data are missing at random (Laird & Ware, 1982; Shah et al., 1997; Andersson & Perlman, 2001; Fairclough et al., 1998; Little & Rubin, 1987; Troxel, Fairclough, Curran & Hahn, 1998). Although the missing at random assumption enables a fairly straightforward likelihood-based analysis without needing to model the missing process (Little & Rubin, 1987), we will argue in §3.1 that this assumption is rarely realistic for nonmonotone missing data. A recent model discussed by Lin et al. (2003) and van der Laan & Robins (2003, Ch. 6) relies on a more plausible assumption about the missingness process which nonetheless assumes no selection on unobservables for the marginal distribution of the responses. In §3.4 we show that it can be viewed as a special case of the model presented in this paper, in which selection bias is absent.

Several proposals also exist for nonmonotone missing data where selection depends on unobservables. With the exception of the selection bias permutation missingness model of Robins et al. (1999), none of the available models is nonparametric identified. The currently available models rely on parametric assumptions for both the full data and the missing data mechanisms (Deltour et al., 1999; Albert, 2000; Ibrahim et al., 2001; Fairclough et al., 1998; Troxel, Fairclough, Curran & Hahn, 1998; Troxel, Lipsitz & Harrington, 1998) or on parametric assumptions for just the missing data process (Rotnitzky et al., 1998; Robins et al., 1995). The selection bias permutation missingness model generalizes the permutation missingness model of Robins (1997) and the sequential coarsening model of Gill & Robins (1997). This model differs from but is related to the nonparametric identified model we propose in the current paper. The two models are compared in §3.5.

2 The formal setting

Consider a longitudinal study design that calls for measurements on a vector of variables L_{it} to be recorded at study cycles $t = 0, \ldots, T$, for the ith of n independent subjects. The vector L_{it}, $t = 1, \ldots, T$, includes an outcome of interest Y_{it} as well as other variables V_{it} recorded for secondary analyses. The vector L_{i0} may include, in addition to Y_{i0} and V_{i0}, a baseline covariate vector X_i of interest.

Suppose that $L_i = (L_{i0}, \ldots, L_{iT})$ is not always fully recorded because some subjects miss some study cycles. In particular, for each t, L_{it} is either completely observed or completely missing and L_{i0} is always observed. Thus, for each t, the observed data for subject i is the vector $O_{it} = (R_{it}, c(R_{it}, L_{it}))$, where R_{it} is a response indicator which equals 1 if L_{it} is observed and is 0 otherwise, and where, for any random vector W, $c(1, W) = W$ and $c(0,$

W) is set to zero by convention. Under our setting, the observed data $O_i = (L_{i0}, O_{i1}, \ldots, O_{iT})$, $i = 1, \ldots, n$, can be regarded as n independent realizations of the random vector $O = (O_0, O_1, \ldots, O_T)$. Here and throughout O_0 denotes L_0. Furthermore, for any vector $Z = (Z_0, \ldots, Z_T)$, $\bar{Z}_t$ denotes the history $(Z_0, \ldots, Z_t)$ up to and including cycle t and $Z_{(t)}$ denotes the vector $(Z_t, \ldots, Z_T)$. Throughout we assume for each t that Y_t is either a continuous or discrete random variable; for any random vector W, $f(Y_t|W)$ denotes a fixed version of the conditional density of Y_t given W with respect to either Lebesgue measure or a counting measure and $\text{pr}(R_t = 1|W)$ denotes a fixed version of the conditional probability that R_t equals 1 given W.

We assume that the nonresponse patterns are nonmonotone so that $R_t = 0$ does not imply that $R_{t+1} = 0$. We additionally assume that no recorded past $\bar{O}_{t-1}$ and no current outcome Y_t can prevent the possibility of returning to study cycle t; that is,

$$\text{pr}(R_t=1|\bar{O}_{t-1}, Y_t)>0 \quad \text{with} \quad \text{probability 1.} \qquad (1)$$

Note that (1) would not hold in a study design where patients are withdrawn when they miss, say, four consecutive visits (Zeuzem et al., 2000) or when all individuals with extreme values of Y_t are so physically impaired that clinic attendance at visit t is impossible. The methods we propose here are not applicable in such cases.

Condition (1) was also assumed in Lin et al. (2003) and van der Laan & Robins (2003, Ch. 6) except that visits could be in continuous rather than in discrete times. These authors obtained identifiability by imposing the additional assumption of sequential explainability,

$$\text{pr}(R_t=1|\bar{O}_{t-1}, Y_t)=\text{pr}(R_t=1|\bar{O}_{t-1}), \qquad (2)$$

which we discuss in detail in §3.4. A conflict as to the number and type of variables to include in the components V_t of the full data vector L_t at each cycle t arises when one wishes to impose both assumptions (1) and (2): to make sequential explainability (2) plausible one would generally wish to choose V_{t-1} to be high-dimensional; however, the positivity assumption (1) may be unrealistic for high-dimensional V_{t-1}, as V_{t-1} may then well include covariates, such as for example the subject's state of consciousness, certain values of which, e.g. being unconscious, preclude the possibility of being observed at occasion t. Since it will often be unrealistic to impose both (2) and (1), we propose to relax the assumption of sequential explainability: we will describe methods for conducting inference about the marginal mean $E(Y_t)$, $t = 1, \ldots, T$, and the conditional mean given baseline covariates $E(Y_t|X)$, $t = 1, \ldots, T$, when (1) holds but (2) may fail.

3 Identifying assumptions

3.1 Preamble

Unless $R_t = 1$ with probability 1, the observed data O identify neither the distribution of Y_t nor its conditional distribution given X because, as Theorem 1 below implies, many distinct conditional laws of Y_t given $\bar{O}_{t-1}$ are compatible with the observed data law. To identify

these distributions we must make unverifiable assumptions. We now review one popular such assumption, that the data are missing at random, and argue that it represents processes that are unlikely to generate nonmonotone missing data patterns in longitudinal studies. We then propose a class of unverifiable assumptions that are naturally suited to conducting sensitivity analysis of the investigator's a priori belief about the process that generates the intermittent nonresponse in a follow-up study. In subsequent sections we discuss inference under any such assumption.

3.2 Missing at random

Robins & Rotnitzky (1992) and Gill et al. (1997) showed that the distribution of the full data vector L is identified if the data are missing at random, provided that there is a positive conditional probability of observing the full data, i.e. $\text{pr}(\bar{R}_T = 1^* | L) > 0$ with probability 1, where 1* denotes the $T \times 1$ vector of ones. The missing at random assumption states that

$$\text{pr}(\bar{R}_T = \bar{r}_T | L) = \text{pr}(\bar{R}_T = \bar{r}_T | L_{(\bar{r}_T)}), \quad (3)$$

where $L_{(\bar{r}_T)}$ denotes the observed part of L when $\bar{R}_T = \bar{r}_T$.

Under any model in which $\text{pr}(\bar{R}_T | L)$ and $f(L)$ are variation independent and the missing at random condition (3) is imposed, the likelihood factorizes into a part that depends on $f(L)$ and another that depends on $\text{pr}(\bar{R}_T | L)$. Any method that obeys the likelihood principle then yields the same inference whether $\text{pr}(\bar{R}_T | L)$ is fully known, unknown or known to follow a model. As a result of this, many authors have proposed analyzing nonmonotone missing data using likelihood-based methods under models that assume missing at random. However, convenient as the missing at random assumption may be, the assumption should only be adopted if it is believed plausible. Following Robins & Gill (1997) we will now argue that missing at random mechanisms that could plausibly generate the observed data in follow-up studies with nonmonotone nonresponse are quite restrictive and would rarely be plausible.

Robins & Gill (1997) showed that the set of missingness probabilities $\text{pr}(\bar{R}_T | L)$ that satisfy missing at random can be divided into two disjoint subsets. The first set contains processes in which the observed data are generated as follows. The variable L_0 is always observed. Then, with probability p_{00} possibly depending on L_0, no further variable is observed. Otherwise one selects which of $L_1, \ldots, L_T$ to observe next by flipping a T-sided coin with probabilities $p_{01}, \ldots, p_{0T}$ that may depend on L_0. One then observes nothing else with probability p_{10} that may depend on L_0 and the most recently observed L_t. Otherwise one selects which of the $T - 1$ still unobserved L_t's to observe next with probabilities $p_{11}, \ldots, p_{1(T-1)}$ that may depend on the already observed L_t's, and so on. The second subset contains all remaining missingness processes that satisfy (3). Robins & Gill (1997) showed that the second subset is not empty. They also showed that to generate the observed data O according to a missingness mechanism in the second set it is required, in the course of the data-generation procedure, to use information about the components of L that are not in $L_{(\bar{R}_T)}$, and thus are ultimately missing, in a subtle and often highly contrived manner to ensure that missing at random holds. In agreement with the discussion in Robins & Gill (1997), we

believe that such missing at random missingness mechanisms are often unrealistic. Consequently, the most reasonable missing at random processes with nonmonotone data are in the first set. Moreover, of the processes in the first set, the only plausible choices for longitudinal data are those in which, with conditional probability equal to 1, the next variable to be observed comes later in time than any variable already observed, as decisions today cannot affect attendance in the past. However, even these are rather unlikely processes when missingness is nonmonotone because they effectively imply for example that, if a patient chooses today to miss his next two visits and then to return, he will not reassess this decision based on evolving time-dependent covariates associated with the response. This is unlikely, as often the decision to miss a given study cycle is influenced by aspects of the subject's health and psychological status that evolved during earlier missed study cycles.

3.3 Occasion-specific tilted models

Part (ii) of Theorem 1 below establishes that, when (1) holds for a fixed t, $t = 1, \ldots, T$, the distributions $f(Y_t|X)$ and $f(Y_t)$ are identified under the following Assumption 1 which postulates that, among subjects with a given observed past O_{t-1}^{-}, the distribution of Y_t in the nonresponders at cycle t is equal to the distribution of Y_t in the responders at cycle t tilted by a known function.

Assumption 1. If

$$\text{pr}(R_t=0|\bar{O}_{t-1})>0, \quad (4)$$

then

$$f(Y_t|\bar{O}_{t-1}, R_t=0)=f(Y_t|\bar{O}_{t-1}, R_t=1)\frac{\exp\{q_t(\bar{O}_{t-1}, Y_t)\}}{E\left[\exp\{q_t(\bar{O}_{t-1}, Y_t)\}|R_t=1, \bar{O}_{t-1}\right]} \quad (5)$$

for some user-specified, i.e. known, function $q_t(O_{t-1}^{-}, Y_t)$.

For ease of reference in the forthcoming discussion, we use $\mathscr{A}_t(q)$ ($\mathscr{A}(q)$) to denote the model for the full data (L, R_T^{-}) defined by (1) and Assumption 1 for a fixed t, for all $t = 1, \ldots, T$. Note that $\mathscr{A}(q)$ is the intersection of models $\mathscr{A}_t(q)$, $t = 1, \ldots, T$.

By Bayes rule, Assumption 1 is equivalent to

$$\text{pr}(R_t=0|\bar{O}_{t-1}, Y_t)=\text{expit}\{h_t(\bar{O}_{t-1})+q_t(\bar{O}_{t-1}, Y_t)\} \quad (6)$$

whenever $f(Y_t|O_{t-1}^{-}, R_t = 1) > 0$,, where expit(.) = exp(.)/{1 + exp(.)} and

$$\exp\{h_t(\bar{O}_{t-1})\}=\frac{\text{pr}(R_t=0|\bar{O}_{t-1})}{\text{pr}(R_t=1|\bar{O}_{t-1})E\left[\exp\{q_t(\bar{O}_{t-1}, Y_t)\}|R_t=1, \bar{O}_{t-1}\right]}. \quad (7)$$

From (6), we interpret each function $q_t(\bar{O}_{t-1}, Y_t)$ as quantifying, on the logistic scale, the magnitude of the residual association between the missingness probability at cycle t and the possibly missing outcome Y_t, after adjustment for the observed past $\bar{O}_{t-1}$. Thus, model $\mathscr{A}(q)$ encodes the investigator's a priori belief of the degree to which, for each t the decision to return to study cycle t is influenced by prognostic factors for Y_t other than those included in the observed past, $\bar{O}_{t-1}$. For example, the choice $q_t(\bar{O}_{t-1}, Y_t) = (1 - R_{t-1})\,\lambda\, Y_t$ encodes the belief that, for those that did not miss the prior cycle $t - 1$, the recorded variables L_{t-1} at the prior cycle together with the observed past $\bar{O}_{t-2}$ prior to cycle $t - 1$ are sufficient to explain missingness at cycle t, but, for those that missed cycle $t - 1$, the observed past $\bar{O}_{t-2}$ is not sufficient to explain missingness at cycle t. The choice $q_t(\bar{O}_{t-1}, Y_t) = \{(1 - R_{t-1})\,\lambda_1 + \lambda_2\}\; Y_t$ additionally allows a residual dependence on the current outcome. In line with the terminology used in some of the missing data literature, we call $q_t(.)$ a selection bias function (Scharfstein et al., 1999).

Part (i) of Theorem 1 below establishes that model $\mathscr{A}(q)$, and therefore $\mathscr{A}_t(q)$ for each t, places no restriction on the observed data law beyond the restriction that $\text{pr}(R_t = 1|\bar{O}_{t-1}) > 0$ for all t. As such, each choice of selection bias functions $q_t(\bar{O}_{t-1}, Y_t)$, $t = 1, \ldots, T$, fits the data perfectly and cannot be rejected by any statistical test. Since there will never be any evidence from the data that can help determine the functions $q_t(\bar{O}_{t-1}, Y_t)$, the analyst should be reluctant to analyze the data solely under one choice of functions $q_t(\bar{O}_{t-1}, Y_t)$. Instead, he should pose a range of plausible selection bias functions and, as a form of sensitivity analysis to his prior beliefs about the missingness mechanism, repeat the analysis under each choice of $q_t(\bar{O}_{t-1}, Y_t)$. This raises the question of how to choose the selection bias functions in practice. We suggest that one chooses, as in the example above, a collection of simple selection bias functions indexed by one or two parameters that are to be varied in a sensitivity analysis. Ideally, the parameterization should satisfy the following properties: it is easily interpretable so that a plausible parameter range can be specified by subject matter experts; values of the parameters equal to zero correspond to the assumption of no selection bias for the outcomes; and nonparametric bounds are attained when the parameters go to $\pm\infty$.

In the following theorem, proved in the Appendix, and throughout, if (4) holds we define

$$m_t\left(\bar{O}_{t-1}\right) \equiv E\left[Y_t \exp\left\{q_t\left(\bar{O}_{t-1}, Y_t\right)\right\} | R_t{=}1, \bar{O}_{t-1}\right] / E\left[\exp\left\{q_t\left(\bar{O}_{t-1}, Y_t\right)\right\} | R_t{=}1, \bar{O}_{t-1}\right].$$

Furthermore, we define $\pi_t(\bar{O}_{t-1}, Y_t) \equiv 1$ if (4) does not hold and

$$\pi_t(\bar{O}_{t-1}, Y_t) \equiv \left[1 + \exp\left\{h_t(\bar{O}_{t-1}) + q_t(\bar{O}_{t-1}, Y_t)\right\}\right]^{-1}$$

otherwise, where $h_t(\bar{O}_{t-1})$ satisfies (7).

Theorem 1. *(i) Model $\mathscr{A}(q)$ determines a model for the law of the observed data whose only restriction is pr$(R_t = 1|\bar{O}_{t-1}) > 0$ with probability 1, for $t = 1, \ldots, T$.*

(ii) The conditional density $f(Y_t|\bar{O}_{t-1})$ *is identified under model* $\mathscr{A}_t(q)$*, and therefore under* $\mathscr{A}(q)$*. Furthermore, under models* $\mathscr{A}_t(q)$ *and* $\mathscr{A}(q)$*,* $E(Y_t)$ *is equal to the following functional of the observed data distribution:*

$$E(Y_t)=E\left\{\frac{R_tY_t}{\pi_t(\bar{O}_{t-1},Y_t)}\right\} \quad (8)$$

$$=E\{R_tY_t+(1-R_t)\,m_t\left(\bar{O}_{t-1}\right)\}. \quad (9)$$

Part (ii) of the Theorem implies the identifiability of the marginal density of Y_t and its conditional distribution given X under model $\mathscr{A}_t(q)$ and therefore also under model $\mathscr{A}(q)$. However, the theorem says nothing about the identifiability of the dependence among outcomes at different occasions. This is so because this dependence is generally not identified under model $\mathscr{A}(q)$. In particular, under model $\mathscr{A}(q)$ the correlations among the repeated outcomes are not identified for any choice of selection-bias function.

3.4 Sequential explainability

The choice $q_t = 0$ postulates the conditional independence of Y_t and R_t given $\bar{O}_{t-1}$. This assumption would hold if the observed past variables $\bar{O}_{t-1}$ included all the predictors of Y_t that explain missingness at cycle t. We therefore refer to the assumption that $q_t = 0$ for all t as the assumption of sequential explainability. Lin et al. (2003) and van der Laan & Robins (2003) consider such processes except that visits occur in continuous time.

The assumption that $q_t = 0$ is less restrictive than the missing at random condition

$$\text{pr}(R_t=1|\bar{R}_{t-1},L)=\text{pr}(R_t=1|\bar{O}_{t-1}) \quad (10)$$

for $t = 1, \ldots, T$, because (10) implies (6) with $q_t = 0$ but the opposite is false. For example, (10) postulates the conditional independence given $\bar{O}_{t-1}$ of R_t with the current components L_t, the future components $L_{(t+1)}$ and the components of $\bar{L}_{t-1}$ corresponding to missed cycles. However, (6) says nothing about the dependence of R_t on future components $L_{(t+1)}$ and the components of $\bar{L}_{t-1}$ corresponding to missed cycles. In fact, while assumption (10) imposes restrictions on the observed data distribution and hence is a testable assumption, this is not true for (6) by part (i) of Theorem 1. Note also that, under model $\mathscr{A}(q)$ with $q_t = 0$ for all t, nonresponse is nonignorable for inference about $f(Y_t)$, $t = 1, \ldots, T$, in the sense that likelihood-based methods do not result in the same inference if $\text{pr}(R_t = 1|\bar{O}_{t-1})$ is known, unknown or known to follow a model. This is because the likelihood of the observed data at each cycle t does not factorize into a part that depends on $\text{pr}(R_t = 1|\bar{O}_{t-1})$ and another that depends on $f(Y_t)$.

3.5 Relationship between model $\mathscr{A}(q)$ and the selection bias permutation missingness model

Members of the class of selection bias permutation missingness models of Robins et al. (1999) are indexed by permutations of the visit subscripts 1, …, T. One such model is particularly appropriate for longitudinal studies and is defined by the assumptions that, for each t,

$$\text{pr}(R_t=1|\bar{O}_{t-1}, Y_t, Y_{(t+1)})>0 \text{ with probability } 1,$$

$$f(Y_t|\bar{O}_{t-1}, Y_{(t+1)}, R_t=0)=\frac{f(Y_t|\bar{O}_{t-1}, Y_{(t+1)}, R_t=1)\ \exp\{q_t(\bar{O}_{t-1}, Y_t, Y_{(t+1)})\}}{E\left[\exp\{q_t(\bar{O}_{t-1}, Y_t, Y_{(t+1)})\}|R_t=1, \bar{O}_{t-1}, Y_{(t+1)}\right]},$$

with $q_t(\bar{O}_{t-1}, Y_t, Y_{(t+1)})$ known. It differs from model $\mathscr{A}(q)$ only in that the future outcome history $Y_{(t+1)}$ is added to each conditioning event and to the function $q_t(\bar{O}_{t-1}, Y_t, Y_{(t+1)})$. Robins et al. (1999) showed that for each q_t, $t = 1, \ldots, T$, this model places no restriction on the distribution of the observed data and identifies the joint distribution of $\bar{Y}_T = (Y_1, \ldots, Y_T)$. Thus this model can be used instead of model $\mathscr{A}(q)$ when the substantive question at issue depends on the joint law of $\bar{Y}_T$ rather than simply on the marginals Y_t of $\bar{Y}_T$. However, the ability to make inferences about the joint law comes at a price as it is more difficult to model $f(Y_t|\bar{O}_{t-1}, Y_{(t+1)}, R_t = 0)$ than $f(Y_t|\bar{O}_{t-1}, R_t = 0)$ both because the conditioning event $(\bar{O}_{t-1}, Y_{(t+1)}, R_t = 0)$ is of greater dimension than the event $(\bar{O}_{t-1}, R_t = 0)$ and because it is less natural to model the law of Y_t given the observed past $\bar{O}_{t-1}$ and the, possibly unobserved, future $Y_{(t+1)}$ than to model the law of Y_t given only the observed past $\bar{O}_{t-1}$.

4 Estimation of the unconditional occasion-specific outcome means

4.1 Nonparametric inference under model $\mathscr{A}(q)$

We now discuss inference about the marginal means $\beta_t^* \equiv E(Y_t)$ under models that assume (6) for user-specified functions $q_t(\bar{O}_{t-1}, Y_t)$, $t = 1, \ldots, T$.

Models $\mathscr{A}(q)$ and $\mathscr{A}_t(q)$ define the same model, i.e. the nonparametric model, for the observed data distribution. Furthermore, as established in part (ii) of Theorem 1, under either model, β_t^* is the same, unique, functional of the observed data law. Thus, inference about β_t^* under either model is identical. In particular, the nonparametric maximum likelihood estimator $\hat{\beta}_{\text{NPML},t}$ of β_t^* under either model is obtained by calculating the expressions in the right-hand side of (8) or (9) under the empirical distribution of O, i.e.

$$\hat{\beta}_{\text{NPML},t}=E_n\left(R_tY_t\left[1+\exp\left\{\hat{h}(\bar{O}_{t-1})_{\text{NPML},t}\right\}\exp\{q_t(\bar{O}_{t-1}, Y_t)\}\right]\right)=E_n\left\{R_tY_t+(1-R_t)\,\hat{m}_{\text{NPML},t}\left(\bar{O}_{t-1}\right)\right\},$$

where $\hat{h}(\bar{O}_{t-1})_{\text{NPML},t} = -\infty$ if $E_n(R_t|\bar{O}_{t-1}) = 1$ and otherwise

$$\hat{h}(\bar{O}_{t-1})_{\mathrm{NPML},t}=\log\left\{\frac{1-E_n(R_t|(\bar{O}_{t-1})}{E_n[R_t\ \exp\{q_t(\bar{O}_{t-1},Y_t)\}|\bar{O}_{t-1}]}\right\},$$

$$\hat{m}_{\mathrm{NPML},t}\ (\bar{O}_{t-1})=\frac{E_n\ \left[Y_t\exp\{q_t(\bar{O}_{t-1},Y_t)\}|R_t=1,\bar{O}_{t-1}\right]}{E_n\ \left[\exp\{q_t(\bar{O}_{t-1},Y_t)\}|R_t=1,\bar{O}_{t-1}\right]},$$

and where for random variables W and Z,

$E_n\ (W|Z)=\sum_{i=1}^n W_iI(Z_i=Z)/\sum_{i=1}^n I(Z_i=Z)$ and $E_n\ (W)=n^{-1}\sum_{i=1}^n W_i$.

Unfortunately, unless T is small and L_t is discrete with few levels, with the sample sizes found in practice the data available for estimating the required conditional expectations will be sparse and consequently the estimator $\hat{\beta}_{\mathrm{NPML},t}$ will be undefined. One could assume that the required conditional expectations are smooth in $\bar{O}_{t-1}$ and use multivariate smoothing techniques to estimate them. However, when $\bar{O}_{t-1}$ is high-dimensional, they would not be well estimated with moderate sample sizes because no two units would have values of $\bar{O}_{t-1}$ close enough to allow the borrowing of information needed for smoothing. Thus, in practice, because of the curse of dimensionality, we are forced to place more stringent dimension-reducing modelling restrictions on the law of the observed data.

Two dimension-reducing strategies are suggested by expressions (8) and (9) for β_t^*. The first strategy is to assume that the function $h_t(\bar{O}_{t-1})$ follows a parametric model,

$$h_t(\bar{O}_{t-1})=h_t(\bar{O}_{t-1};\alpha_t^*), \quad (11)$$

where $h_t(\bar{O}_{t-1};\alpha_t)$ is a known function smooth in α_t, and α_t^* is an unknown $p_{t,h}\times 1$ parameter vector. The second strategy is to assume that $m_t(\bar{O}_{t-1})$ follows a parametric model,

$$m_t\ (\bar{O}_{t-1})=m_t\ (\bar{O}_{t-1};\theta_t^*), \quad (12)$$

where $m_t(\bar{O}_{t-1};\theta_t)$ is a known function, smooth in θ_t, and θ_t^* is an unknown $p_{t,m}\times 1$ parameter vector.

We use $\mathscr{B}_t(q)$ ($\mathscr{B}(q)$) to denote the model for the full data $(L,\bar{R}_T)$ defined by the assumptions of model $\mathscr{A}_t(q)$ and the additional restrictions (4) and (11) specified just for a fixed t (for all t). Likewise, we use $\mathscr{C}_t(q)$ ($\mathscr{C}(q)$) to denote the model for the full data $(L,\bar{R}_T)$ defined by the assumptions of model $\mathscr{A}_t(q)$ and the additional restriction (12) specified just for a fixed t, for all t.

Models (11) and (12) are not in themselves of scientific interest. However, in practice we are forced to impose one of the two models. Estimation of β_t^* under model $\mathscr{B}_t(q)$ is not entirely satisfactory because the resulting estimators of β_t^* can be biased if model (11) is incorrect, and estimation under $\mathscr{C}_t(q)$ suffers from the same limitation if model (12) is incorrect. Luckily there is an alternative strategy for estimation of β_t^*. This consists of computing an estimator that is consistent and asymptotically normal in the union model $\mathscr{B}_t(q) \cup \mathscr{C}_t(q)$, i.e. an estimator of β_t^* that is consistent and asymptotically normal so long as one of the models $\mathscr{B}_t(q)$ or $\mathscr{C}_t(q)$, but not necessarily both, is correctly specified. Following Robins (2000) and Robins & Rotnitzky (2001) we call such an estimator a doubly robust estimator in the union model $\mathscr{B}_t(q) \cup \mathscr{C}_t(q)$ as it can protect against misspecification of either (11) or (12), although not against simultaneous misspecification of both. The following definition introduces a generalization of double robustness.

Definition. *Given a collection* $\{\mathscr{M}_u; u \in \mathbb{U}$ *of models for a law F indexed by the elements of a finite set* $\mathbb{U}$ *with K elements, we say that an estimator* $\hat{\lambda}$ *of a parameter* $\lambda \equiv \lambda(F)$ *is a K-multiply robust estimator in the union model* $\cup_{u\in\mathbb{U}}\mathscr{M}_u$ *if it is a consistent and asymptotically normal estimator of* λ *when one of the models* $\mathscr{M}_u$, $u \in \mathbb{U}$ *but not necessarily more than one of them, holds.*

4.2 Doubly and 2^T-multiply robust estimation

In this subsection we propose a doubly robust estimator of β_t^* in the union model $\mathscr{B}_t(q) \cup \mathscr{C}_t(q)$ and a multiply robust estimator of β* in the union model $\cup_{u\in\mathbb{U}}\mathscr{M}_u(q)$. Throughout, $\mathbb{U}$ denotes the collection of $T\times 1$ vectors u whose components are either 0 or 1. For each such vector u, $\mathscr{M}_u = \{\cap_{t:u_t=0}\mathscr{B}_t(q)\} \cap \{\cap_{t:u_t=1}\mathscr{C}_t(q)\}$. Thus, a 2^T-multiply-robust estimator of β* is consistent and asymptotically normal for $\beta^* = (\beta_1^*, \ldots, \beta_T^*)$ so long as at each t one of the models $\mathscr{B}_t(q)$ or $\mathscr{C}_t(q)$, but not necessarily both, is correctly specified.

To construct the doubly and 2^T-multiply robust estimators we reason as follows. Suppose that we have specified working models (11) and (12). For any constant column vector d_t and conformable column vector functions $\varphi_t(\bar{O}_{t-1})$ and $\psi_t(\bar{O}_{t-1})$, define

$$H_t(d_t, \phi_t, \beta_t, \alpha_t) = \frac{d_t R_t \varepsilon_t(\beta_t)}{\pi_t(\bar{O}_{t-1}, Y_t; \alpha_t)} + \left\{1 - \frac{R_t}{\pi_t(\bar{O}_{t-1}, Y_t; \alpha_t)}\right\} \phi_t(\bar{O}_{t-1}) \tag{13}$$

$$M_t(d_t, \psi_t, \beta_t, \theta_t) = d_t R_t \varepsilon_t(\beta_t) + d_t (1 - R_t) \left\{m_t(\bar{O}_{t-1}; \theta_t) - \beta_t\right\} + R_t \psi_t(\bar{O}_{t-1}) \exp\{q_t(\bar{O}_{t-1}, Y_t)\} \left\{Y_t - m_t(\bar{O}_{t-1}; \theta_t)\right\}, \tag{14}$$

where $\varepsilon_t(\beta_t) = Y_t - \beta_t$ and $\pi_t(\bar{O}_{t-1}, Y_t; \alpha_t) \equiv [1 + \exp\{h_t(\bar{O}_{t-1}; \alpha_t) + q_t(\bar{O}_{t-1}, Y_t)\}]^{-1}$. In the Appendix we show that, provided we choose $\varphi_t(\bar{O}_{t-1})$ in (13) and $\psi_t(\bar{O}_{t-1})$ in (14) to have the specific functional forms given by

$$\phi_{\theta_t,\beta_t,t}(\bar{O}_{t-1})=m_t\left(\bar{O}_{t-1};\theta_t\right)-\beta_t,\psi_{\alpha_t,t}\left(\bar{O}_{t-1}\right)=\exp\left\{h_t(\bar{O}_{t-1};\alpha_t)\right\}, \quad (15)$$

for any fixed θ_t, α_t and β_t, $H_t(1, \varphi_{\theta_t\beta_t t}, \beta_t, \alpha_t)$ and $M_t(1, \psi_{\alpha_t t}, \beta_t, \theta_t)$ are identical, where 1 is a scalar constant function equal to 1. We therefore write, for short,

$$Q_t\left(\beta_t,\theta_t,\alpha_t\right)\equiv H_t(1,\phi_{\theta_t,\beta_t,t},\beta_t,\alpha_t)=M_t(1,\psi_{\alpha_t,t},\beta_t,\theta_t).$$

In the Appendix we also show that if model (11) holds then, for any θ_t and regardless of whether or not model (12) holds, $H_t(1,\phi_{\theta_t,\beta_t^*,t},\beta_t^*,\alpha_t^*)$ has mean zero and therefore so does $Q_t\left(\beta_t^*,\theta_t,\alpha_t^*\right)$. Furthermore, if model (12) holds then, for any α_t and regardless of whether or not model (11) holds, $M_t(1,\psi_{\alpha_t,t},\beta_t^*,\theta_t^*)$ has mean zero and therefore so does $Q_t(\beta_t^*,\theta_t^*,\alpha_t)$. These results suggest that we can construct a doubly robust estimator $\hat{\beta}_t$ of β_t^*, throughout also denoted by $\hat{\beta}_t(\psi_t, \varphi_t)$, in model $\mathscr{B}_t(q)\cup\mathscr{C}_t(q)$ by solving the scalar estimating equation

$$E_n\left\{Q_t\left(\beta_t,\hat{\theta}_t,\hat{\alpha}_t\right)\right\}=0, \quad (16)$$

where $\hat{\theta}_t$ and $\hat{\alpha}_t$ solve

$$E_n\left\{M_t(0,\psi_t,0,\theta_t)\right\}=0, E_n\left\{H_t(0,\phi_t,0,\alpha_t)\right\}=0,$$

using arbitrary $p_{t,m}\times 1$ and $p_{t,h}\times 1$ functions $\psi_t(\bar{O}_{t-1})$ and $\varphi_t(\bar{O}_{t-1})$ respectively. Theorem 2 below establishes that $\hat{\beta}_t$ is doubly robust in model $\mathscr{B}_t(q)\cup\mathscr{C}_t(q)$ and that $\hat{\beta}=(\hat{\beta}_1,\ldots,\hat{\beta}_T)'$, throughout also denoted by $\hat{\beta}(\psi,\varphi)$, is 2^T-multiply robust in model $\cup_{u\in\mathbb{U}}\mathscr{M}_u(q)$.

To state the asymptotic properties of $\hat{\beta}_t$ and $\hat{\beta}$ in Theorem 2, we define

$$U_t\left(\beta_t,\theta_t,\alpha_t\right)=Q_t\left(\beta_t,\theta_t,\alpha_t\right)-I_{\alpha_t,Q_t}\left(\beta_t,\theta_t,\alpha_t\right)I_{\alpha_t,H_t}^{-1}\left(\alpha_t\right)H_t(0,\phi_t,0,\alpha_t)-I_{\theta_t,Q_t}\left(\beta_t,\theta_t,\alpha_t\right)I_{\theta_t,M_t}^{-1}\left(\theta_t\right)M_t(0,\psi_t,0,\theta_t)$$

$$Q\left(\beta,\theta,\alpha\right)=\left(Q_1\left(\beta_1,\theta_1,\alpha_1\right),\ldots,Q_T\left(\beta_T,\theta_T,\alpha_T\right)\right)'$$

$$U\left(\beta,\theta,\alpha\right)=\left(U_1\left(\beta_1,\theta_1,\alpha_1\right),\ldots,U_T\left(\beta_T,\theta_T,\alpha_T\right)\right)',$$

where for any random vector function $K(\tau,\eta)$ of a parameter (τ,η), $I_{\tau,K}(\tau,\eta)$ denotes $E\{\partial K(\tau,\eta)/\partial\tau\}$ and $\partial K(\tau,\eta)/\partial\tau$ is a derivative matrix with (i,j) entry equal to $\partial K_i(\tau,\eta)/\partial\tau_j$. In what follows and throughout, we use a hat to indicate expectations calculated under the empirical distribution of the observed data and evaluation at $(\beta_t,\theta_t,\alpha_t)=(\hat{\beta}_t,\hat{\theta}_t,\hat{\alpha}_t)$; for example,

$$\hat{U}_t = Q_t\left(\hat{\beta}_t, \hat{\theta}_t, \hat{\alpha}_t\right) - \hat{I}_{\alpha_t, Q_t}\hat{I}^{-1}_{\alpha_t, H_t} H_t(0, \phi_t, 0, \hat{\alpha}_t) - \hat{I}_{\theta_t, Q_t}\hat{I}^{-1}_{\theta_t, M_t} M_t(0, \psi_t, 0, \hat{\theta}_t)$$

and $\hat{U} = (\hat{U}_1, \ldots, \hat{U}_T)'$, where $\hat{I}_{\alpha_t, Q_t} = E_n\{\partial Q_t(\hat{\beta}_t, \hat{\theta}_t, \alpha_t)/\partial\alpha_t|_{\alpha_t = \hat{\alpha}_t}\}$, and so on.

Parts (i) and (iii) of Theorem 2, proved in the Appendix, state the asymptotic distribution of $\hat{\beta}_t$ and $\hat{\beta}$ under models $\mathscr{B}_t(q) \cup \mathscr{C}_t(q)$ and $\cup_{u \in \mathbb{U}} \mathscr{M}_u(q)$, respectively. Parts (ii) and (iv) state that the asymptotic variance of $\hat{\beta}_t$ and $\hat{\beta}$ under these models remains the same regardless of the choice of the functions ψ_t and φ_t used to compute the estimators $\hat{\theta}_t$ and $\hat{\alpha}_t$ of θ_t^* and α_t^*, when in fact the true data-generating process satisfies models $\mathscr{B}_t(q) \cap \mathscr{C}_t(q)$ and $\cap_{t=1}^T\{\mathscr{B}_t(q) \cap \mathscr{C}_t(q)\}$, respectively. In practice, the choice of functions ψ_t and φ_t should therefore have little impact on the efficiency of $\hat{\beta}_t$ and $\hat{\beta}$ when the models $\mathscr{B}_t(q)$ and $\mathscr{C}_t(q)$, $t = 1, \ldots, T$, cannot be rejected using efficient goodness-of-fit tests. In what follows, for any matrix A, $A^{\otimes 2}$ denotes AA'.

Theorem 2. *Suppose that the regularity conditions stated in the* Appendix *hold.*

(i) Under model $\mathscr{B}_t(q) \cup \mathscr{C}_t(q)$, $\sqrt{n}\left(\hat{\beta}_t - \beta_t^*\right) \to N(0, \Gamma_t)$ *in distribution, where*

$$\Gamma_t = E\left[\{I^{-1}_{\beta_t, Q_t}(\beta_t^*, \theta_t^0, \alpha_t^0)\, U_t(\beta_t^*, \theta_t^0, \alpha_t^0)\}^{\otimes 2}\right]$$

and θ_t^0 and α_t^0 *are the probability limits of* $\hat{\theta}_t$ and $\hat{\alpha}_t$. *The matrix* Γ_t *can be consistently estimated with*

$$\hat{\Gamma}_t = \hat{I}^{-1}_{\beta_t, Q_t} E_n\left(\hat{U}_t^{\otimes 2}\right) \hat{I}^{-1\prime}_{\beta_t, Q_t}.$$

(ii) Let $\left(\psi_t^{(1)}, \phi_t^{(1)}\right)$ and $\left(\psi_t^{(2)}, \phi_t^{(2)}\right)$ *be two pairs of distinct* $p_{t,m} \times 1$ *and* $p_{t,h} \times 1$ functions (ψ_t, φ_t) of O_{t-1}. Then, under the intersection model $\mathscr{B}_t(q) \cap \mathscr{C}_t(q)$,
$\sqrt{n}\left\{\hat{\beta}_t\left(\psi_t^{(1)}, \phi_t^{(1)}\right) - \hat{\beta}_t\left(\psi_t^{(2)}, \phi_t^{(2)}\right)\right\} = o_p(1)$.

(iii) Under model $\cup_{u \in \mathbb{U}} \mathscr{M}_u(q)$, $\sqrt{n}(\hat{\beta} - \beta^*) \to N(0, \Gamma)$ *in distribution, where*

$$\Gamma = E\left[\{I^{-1}_{\beta, Q}(\beta^*, \theta^0, \alpha^0)\, U(\beta^*, \theta^0, \alpha^0)\}^{\otimes 2}\right]$$

and θ^0 *and* α^0 *are the probability limits of* $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_T)'$ *and* $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_T)'$. *The matrix* Γ *can be consistently estimated with* $\hat{\Gamma} = \hat{I}^{-1}_{\beta, Q} E_n\left(\hat{U}^{\otimes 2}\right) \hat{I}^{-1\prime}_{\beta, Q}$.

(iv) Let $(\psi^{(1)}, \varphi^{(1)})$ *and* $(\psi^{(2)}, \varphi^{(2)})$ *be two distinct sets of functions* $\{(\psi_t(O_{t-1}), \varphi_t(O_{t-1})), t = 1, \ldots, T\}$. *Then, under the intersection model*
$\cap_{t=1}^T\{\mathscr{B}_t(q) \cap \mathscr{C}_t(q)\}$, $\sqrt{n}\left\{\hat{\beta}\left(\psi^{(1)}, \phi^{(1)}\right) - \hat{\beta}\left(\psi^{(2)}, \phi^{(2)}\right)\right\} = o_p(1)$.

It is possible to show that, at the intersection model $\mathscr{B}_t(q) \cap \mathscr{C}_t(q)$, every doubly robust estimator $\hat{\beta}_t(\psi_t, \varphi_t)$ has asymptotic variance that attains neither the semiparametric variance bound for estimation of β_t^* under model $\mathscr{C}_t(q)$ nor, except when $q_t = 0$, the semiparametric variance bound for estimation of β_t under model $\mathscr{B}_t(q)$. In our opinion, the hope to control bias is more important than efficiency concerns, and we therefore recommend using doubly or 2^T-multiply robust estimators of β_t^* and β*, respectively.

So far, we have not allowed the parameters α_t in model (11) and θ_t in model (12), respectively, to be shared across occasions. When we are faced with sample sizes that are not large enough to yield well-behaved estimators of α_t^* and θ_t^* at each occasion *t*, two dimension-reducing strategies can be envisaged. The first strategy is to reduce further the dimension of models (11) and (12) at each *t*. The second strategy is to allow parameters α_t and θ_t, respectively, to be shared across occasions and to compute an estimator of β* that is consistent and asymptotically normal so long as at least one of $\cap_{t=1}^T \mathscr{B}_t(q)$ or $\cap_{t=1}^T \mathscr{C}_t(q)$ holds. Denote by α and θ the resulting $p_h \times 1$ and $p_m \times 1$ parameter vectors indexing models (11) and (12) respectively, for all occasions *t*. Then such an estimator $\tilde{\beta}$ of β* can be obtained by solving estimating equations (16) at each occasion *t*, in which $\hat{\theta}_t \equiv \hat{\theta}$ and $\hat{\alpha}_t \equiv \hat{\alpha}$ now solve $E_n\{M(\psi, \theta)\} = 0$ with $M(\psi,\theta)=\sum_{t=1}^T M_t(0,\psi_t,0,\theta)$ and $E_n\{H(\varphi, \alpha)\} = 0$ with $H(\phi,\alpha)=\sum_{t=1}^T H_t(0,\phi_t,0,\alpha)$, where ψ and φ are vectors of $p_m \times 1$ functions $\psi_t(\bar{O}_{t-1})$ and $p_h \times 1$ functions $\varphi_t(\bar{O}_{t-1})$, $t = 1, \ldots, T$, respectively. Parts (iii) and (iv) of Theorem 2 continue to hold for the resulting estimator $\tilde{\beta}$ if we replace model $\cup_{u\in\mathbb{U}}\mathscr{M}_u(q)$ by $\{\cap_{t=1}^T \mathscr{B}_t(q)\} \cup \{\cap_{t=1}^T \mathscr{C}_t(q)\}$, $\hat{\beta}$ by $\tilde{\beta}$ and $U(\beta, \theta, \alpha)$ by

$$Q(\beta,\theta,\alpha) - I_{\alpha,Q}(\beta,\theta,\alpha)\, I_{\alpha,H}^{-1}(\alpha)\, H(\phi,\alpha) - I_{\theta,Q}(\beta,\theta,\alpha)\, I_{\theta,M}^{-1}(\theta)\, M(\psi,\theta).$$

5 Estimation of the occasion-specific conditional outcome means given baseline covariates

Suppose now that we are interested in inference about a parameter, which we denote again by β*, indexing a regression model for the conditional mean of Y_t, $t = 1, \ldots, T$, given baseline covariates *X*; that is, for $t = 1, \ldots, T$,

$$E(Y_t|X) = g_t(X;\beta^*), \quad (17)$$

where $g_t(X; \beta)$ is a known function that is smooth in β and $\beta^* \in \Theta \subseteq \mathbb{R}^r$ is unknown. Denote by $\mathscr{A}^*(q)$ the model for $(L, \bar{R}_T)$ defined by the restrictions of model $\mathscr{A}(q)$ and the additional restriction (17) for all $t = 1, \ldots, T$.

Part (i) of Theorem 1 is no longer true if model $\mathscr{A}(q)$ is replaced with model $\mathscr{A}^*(q)$, i.e. $\mathscr{A}^*(q)$ does not determine a nonparametric model for the observed data law. Hence, in principle, under (17) the postulated functions q_t may sometimes be subject to an empirical test. Moreover, β* and q_t, $t = 1, \ldots, T$, may be jointly identified under (17). However, there

would generally be very limited independent information about β* and q_t, $t = 1, \ldots, T$, and therefore their joint estimation would require very large sample sizes. In fact, it follows from Proposition B1, Part 6, of Rotnitzky et al. (1998) that, when pr $(R_1 = \ldots = R_T = 0|L) > \sigma > 0$ with probability 1 and both h_t and q_t, $t = 1, \ldots, T$, in (6) are unknown, β* cannot be estimated at rate $\sqrt{n}$. Thus, we continue to recommend that one regard the functions q_t, $t = 1, \ldots, T$, as fixed and known when estimating β* and then vary these functions in a sensitivity analysis.

As was the case for estimation of the marginal means $E(Y_t)$, unless T is small and L_t is discrete with few levels, inference about β* requires placing dimension-reducing assumptions on either h_t or m_t, in addition to the restrictions of model $\mathscr{A}^*(q)$. We therefore consider, for each $t = 1, \ldots, T$, models $\mathscr{B}_t^*(q)$ and $\mathscr{C}_t^*(q)$ defined like models $\mathscr{B}_t(q)$ and $\mathscr{C}_t(q)$ respectively but with the additional restriction (17). Furthermore, we let $\mathscr{M}_u^*(q)$ be defined like $\mathscr{M}_u(q)$ in §4.2 but with $\mathscr{B}_t^*(q)$ and $\mathscr{C}_t^*(q)$ instead of $\mathscr{B}_t(q)$ and $\mathscr{C}_t(q)$ so that the union model $\cup_{u\in\mathbb{U}}\mathscr{M}_u^*(q)$ stands for the model in which at each t either $\mathscr{B}_t^*(q)$ or $\mathscr{C}_t^*(q)$ holds, but not necessarily both. In this section we consider estimation of β* under the union model $\cup_{u\in\mathbb{U}}\mathscr{M}_u^*(q)$.

Although, as shown in the Appendix, the restrictions defining $\mathscr{C}_t(q)$ for each fixed t, and indeed simultaneously for all t, are guaranteed to be compatible, the same is not true for $\mathscr{C}_t^*(q)$. To be specific, for each t, given a function q_t the function $m_t(\bar{O}_{t-1})$ and the conditional mean function $E(Y_t|X)$ are not variation independent; that is, fixing one restricts the range of possible functions for the other. Thus, it may happen that there exists no joint distribution of $(L, \bar{R}_T)$ of which the marginal of O is the observed data distribution and that satisfies simultaneously (5), (12) and (17). Furthermore, even if such incompatibility is not present, it may still happen that the parameter space for β* under $\mathscr{C}_t^*(q)$ is much smaller than that under $\mathscr{A}^*(q)$. This is clearly undesirable because any reasonable dimension-reducing strategy should not, a priori, eliminate values of β* regarded plausible under the model of scientific interest. The following simple example illustrates these points.

Example. Suppose that $T = 1$, $L_0 = X$ and Y_1 is binary. Suppose that in (17) we assume that logit pr $(Y_1 = 1|X) = \beta_0 + \beta_1 X$, $q_1(Y_1, X, V) = \lambda Y_1$ with $\lambda > 0$ and logit pr $(Y_1 = 1|R_1 = 0, X) = \theta_0 + \theta_1 X$. Under this model logit pr $(Y_1 = 1|R_1 = 1, X) = \lambda + \theta_0 + \theta_1 X$ and hence pr $(Y_1 = 1|R_1 = 1, X) >$ pr $(Y_1 = 1|R_1 = 0, X)$. Therefore, since

$$\text{pr}\,(R_1{=}1|X){=}\frac{\text{pr}\,(Y_1{=}1|X)-\text{pr}\,(Y_1{=}1|R_1{=}0,X)}{\text{pr}\,(Y_1{=}1|R_1{=}1,X)-\text{pr}\,(Y_1{=}1|R_1{=}0,X)},$$

it must be that pr $(Y_1 = 1|R_1 = 0, X) <$ pr $(Y_1 = 1|X) <$ pr $(Y_1 = 1|R_1 = 1, X)$. This implies that logit pr $(Y_1 = 1|X) = \lambda^* + \theta_0 + \theta_1 X$ for some $0 < \lambda^* < \lambda$. In particular, $\beta_1 = \theta_1$. It follows that the parameter space for β may be more restricted once we impose the restrictions on pr $(Y_1 = 1|R_1 = 0, X)$. For example, if the model for pr $(Y_1 = 1|R_1 = 0, X)$ restricts θ_1 to lie in a strict subset of the real line and the model for pr $(Y_1 = 1|X)$ leaves β_1 unrestricted, then, once the model for pr $(Y_1 = 1|R_1 = 0, X)$ is imposed, the parameter space for β_1 is reduced to that

for θ_1. The models would even become incompatible if a probit regression were considered for pr ($Y_1 = 1|X$) and a logistic regression for pr ($Y_1 = 1|R_1 = 0, X$).

In the Appendix we show that $\mathscr{B}_t^*(q)$, and indeed $\cap_{t=1}^T \mathscr{B}_t^*(q)$, impose restrictions that are always compatible.

When model $\mathscr{C}_t^*(q)$ is incompatible with model (17) then an estimator of β* that is consistent under the union model $\mathscr{B}_t^*(q) \cup \mathscr{C}_t^*(q)$ actually converges in probability to β* only if the working model $\mathscr{B}_t^*(q)$ holds and hence the estimator is not really doubly robust. We do not regard this theoretical difficulty to be of concern in practice since it is ameliorated if one postulates a richly parameterized model (12). To see this note that, when no restriction is placed on $m_t(\bar{O}_{t-1})$, model $\mathscr{C}_t^*(q)$ becomes model $\mathscr{A}_t^*(q)$ defined like $\mathscr{A}_t(q)$ but with the additional restriction (17). As shown in the Appendix, the restrictions defining model $\mathscr{A}_t^*(q)$ are always compatible. Consequently, if (17) is correctly specified then a flexible model for $m_t(\bar{O}_{t-1})$ should result in a nearly correctly specified model $\mathscr{C}_t^*(q)$. In order to highlight the possibility of model incompatibility, we refer to estimators that are consistent and asymptotically normal under model $\cup_{u\in\mathbb{U}} \mathscr{M}_u^*(q)$ as generalized 2^T-multiply robust estimators.

In agreement with the discussion in Robins & Rotnitzky (2001), we recommend estimating β* with generalized 2^T-multiply robust estimators because such estimators are expected to have small asymptotic bias if, at each $t = 1, \ldots, T$, at leats one of the models $\mathscr{B}_t^*(q)$ or $\mathscr{C}_t^*(q)$ is approximately correct.

We construct generalized 2^T-multiply robust estimators of β* in model $\cup_{u\in\mathbb{U}} \mathscr{M}_u^*(q)$ as follows. Redefine $H_t(d_t, \varphi_t, \beta_t, \alpha_t)$, $M_t(d_t, \psi_t, \beta_t, \theta_t)$ and $\varphi_{\theta_t\beta_t t}(\bar{O}_{t-1})$ as in (13), (14) and (15) but with $\varepsilon_t(\beta_t)$ replaced by $\varepsilon_t(\beta) = Y_t - g_t(X;\beta)$, $m_t(\bar{O}_{t-1};\theta_t) - \beta_t$ by $m_t(\bar{O}_{t-1};\theta_t) - g_t(X;\beta)$ and with $d_t = d_t(X)$ an arbitrary conformable vector function of X. With these redefinitions, it is true that $H_t(d_t, d_t\varphi_{\theta_t\beta,t}, \beta, \alpha_t) = M_t(d_t, d_t\psi_{\alpha_t,t}, \beta, \theta_t)$ for any arbitrary $r \times 1$ vector function $d_t(X)$. We therefore write

$$Q_t(d_t, \beta, \theta_t, \alpha_t) \equiv H_t(d_t, d_t\phi_{\theta_t,\beta,t}, \beta, \alpha_t) = M_t(d_t, d_t\psi_{\alpha_t,t}, \beta, \theta_t).$$

Similarly to §4.2, we construct generalized 2^T-multiply robust estimators $\hat{\beta}$ of β*, also denoted throughout by $\hat{\beta}(d, d_\theta, \psi, d_\alpha, \varphi)$, in model $\cup_{u\in\mathbb{U}} \mathscr{M}_u^*(q)$ by solving an $r \times 1$ estimating equation of the form

$$E_n\left\{Q\left(d, \beta, \tilde{\theta}, \tilde{\alpha}\right)\right\} = 0, \quad (18)$$

where $Q(d, \beta, \theta, \alpha) = \sum_{t=1}^T Q_t(d_t, \beta, \theta_t, \alpha_t)$, $d(X) = (d_1(X), \ldots, d_T(X))$ for arbitrary $r \times 1$ vector functions $d_t(X)$, $t = 1, \ldots, T$, and where $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_T)'$ and $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_T)'$ solve

$$E_n\{M_t(d_{t\theta},\psi_t,\beta,\theta_t)\}=0,\ E_n\{H_t(d_{t\alpha},\phi_t,\beta,\alpha_t)\}=0$$

respectively, for $t = 1, \ldots, T$, using arbitrary collections of functions,

$$d_\theta=\{d_{t\theta}(X): d_{t\theta}(\cdot) \text{ is a } p_{t,m}\times 1 \text{ vector value map}, t=1,\ldots,T\},$$

$$\psi=\{\psi_t(\bar{O}_{t-1}): \psi_t(\cdot) \text{ is a } p_{t,m}\times 1 \text{ vector value map}, t=1,\ldots,T\},$$

$$d_\alpha=\{d_{t\alpha}(X): d_{t\alpha}(\cdot) \text{ is a } p_{t,h}\times 1 \text{ vector value map}, t=1,\ldots,T\},$$

$$\phi=\{\phi_t(\bar{O}_{t-1}): \phi_t(\cdot) \text{ is a } p_{t,h}\times 1 \text{ vector value map}, t=1,\ldots,T\}.$$

Parts (iii) and (iv) of Theorem 2 remain valid if we replace $\cup_{u\in\mathbb{U}}\mathscr{M}_u(q)$ by $\cup_{u\in\mathbb{U}}\mathscr{M}_u^*(q)$, $\hat{\beta}(\psi^{(j)},\varphi^{(j)})$ by $\hat{\beta}\left(d,d_\theta^{(j)},\psi^{(j)},d_\alpha^{(j)},\phi^{(j)}\right)$, $j = 1, 2$, $Q_t(\beta,\theta_t,\alpha_t)$ by $Q_t(d_t,\beta,\theta_t,\alpha_t)$ and we use the redefinition $U(\beta,\theta,\alpha)=\sum_{t=1}^{T}U_t(\beta,\theta_t,\alpha_t)$, where

$$\begin{aligned}U_t(\beta,\theta_t,\alpha_t)&=Q_t(d_t,\beta,\theta_t,\alpha_t)\\&\quad-I_{\alpha_t,Q_t}(\beta,\theta_t,\alpha_t)\,I^{-1}_{\alpha_t,H_t}(\alpha_t)\,H_t(d_{t\alpha},\phi_t,\beta,\alpha_t)\\&\quad-I_{\theta_t,Q_t}(\beta,\theta_t,\alpha_t)\,I^{-1}_{\theta_t,M_t}(\theta_t)\,M_t(d_{t\theta},\psi_t,\beta,\theta_t),\end{aligned}$$

with $d_t = d_t(X)$, $d_{t\alpha} = d_{t\alpha}(X)$ and $d_{t\theta} = d_{t\theta}(X)$.

Theorem 3 below, proved in the Appendix, provides the optimal $r\times T$ matrix function $d_{\text{opt}}(X) = (d_{1,\text{opt}}(X), \ldots, d_{T,\text{opt}}(X))$ in the sense that, among all estimators $\hat{\beta}(d, d_\theta, \psi, d_\alpha, \varphi)$ that solve (18) using an arbitrary $r\times T$ matrix functions $d(X)$ and fixed collections of functions d_θ, ψ, d_α and φ, the estimator with the smallest asymptotic variance is the one that uses $d = d_{\text{opt}}$. In particular, since under laws in the intersection model $\cap_{t=1}^{T}\{\mathscr{B}_t^*(q)\cap\mathscr{C}_t^*(q)\}$ the limiting distribution of $\hat{\beta}(d, d_\theta, \psi, d_\alpha, \varphi)$ does not depend on the choice of d_θ, ψ, d_α and φ, we conclude that the estimator that solves (18) using $d = d_{\text{opt}}$ has the smallest asymptotic variance among all estimators $\hat{\beta}(d, d_\theta, \psi, d_\alpha, \varphi)$ under any law in $\cap_{t=1}^{T}\{\mathscr{B}_t^*(q)\cap\mathscr{C}_t^*(q)\}$. In the following theorem, $\Gamma(d, d_\theta, \psi, d_\alpha, \varphi)$ denotes the variance of the limiting normal distribution of $\sqrt{n}\{\hat{\beta}(d, d_\theta, \psi, d_\alpha, \varphi) - \beta^*\}$ under model $\cup_{u\in\mathbb{U}}\mathscr{M}_u^*(q)$ and $U^*(\beta,\theta,\alpha)\equiv(U_1^*(\beta,\theta_1,\alpha_1),\ldots,U_T^*(\beta,\theta_T,\alpha_T))'$, where each $U_t^*(\beta,\theta_t,\alpha_t)$ is defined as $U_t(\beta,\theta_t,\alpha_t)$ but with the $r\times 1$ function $d_t(X)$ replaced by the constant real valued function $d_t(X)\equiv 1$. Also, θ^0 and α^0 denote the probability limits of $\tilde{\theta}$

and $\tilde{\alpha}$. In addition, for any pair of conformable square matrices A and B, $A \geq B$ indicates that $A - B$ is positive semidefinite.

Theorem 3. *For every fixed collection of functions* d_θ, ψ, d_α *and* φ, *we have that* $\Gamma(d_{opt}, d\theta, \psi, d_\alpha, \varphi) \leq \Gamma(d, d_\theta, \psi, d_\alpha, \varphi)$, *where*

$$d_{\text{opt}}(X) = E\left\{ \left. \frac{\partial U^*(\beta, \theta^0, \alpha^0)}{\partial \beta} \right|_{\beta=\beta^*} \middle| X \right\} [\text{var}\{U^*(\beta^*, \theta^0, \alpha^0)|X\}]^{-1}.$$

6 Simulation study

We conducted two simulation experiments. The first compares the behaviour in finite samples of the 2^T-multiply robust estimators of marginal means with competitors that are not 2^T-multiply robust, and the second evaluates the behaviour of generalized 2^T-multiply robust estimators of parameters of regression models for the marginal means. Each experiment was based on 1000 replications of random samples of size 500 generated as follows. In both experiments, for $t > 1$, L_t comprised just the outcome Y_t. In the first experiment L_0 was standard normal and, for $t = 1, \ldots, 4$, given $(\bar{L}_{t-1}, R_{t-1})$, R_t was generated from

$$\text{pr}(R_t = 0 | \bar{R}_{t-1}, \bar{L}_{t-1}) = \text{expit}[-2.25 + 0.25t + 0.5\{1 + I(t=1)\}L_0 + 2(1 - R_{t-1}) + 2R_{t-1}\varepsilon_{t-1}],$$

and then L_t was generated as $L_t = 2t + 3L_0 + R_{t-1} + 2R_{t-1}\varepsilon_{t-1} - \gamma R_t + \varepsilon_t$ for the choices $\gamma = 0$ and $\gamma = -0.5$, where $\varepsilon_1, \ldots, \varepsilon_4$ are four independent standard normal variates. It is easy to check that the law of our simulated data satisfies the restrictions of $\mathscr{B}(q)$ and $\mathscr{C}(q)$ with $q_t(\bar{O}_{t-1}, Y_t) = \gamma Y_t$, $h_t(\bar{O}_{t-1}; \alpha_t) = \alpha_{t0} + \alpha_{t1}L_0 + I(t > 1)(\alpha_{t2}R_{t-1} + \alpha_{t3}R_{t-1}Y_{t-1})$ and $m_t(\bar{O}_{t-1}; \theta_t) = \theta_{t0} + \theta_{t1}L_0 + I(t > 1)(\theta_{t2}R_{t-1} + \theta_{t3}R_{t-1}Y_{t-1})$ for specific vector values α_t and θ_t.

In the second experiment, given a standard normal variate X, we generated a 4×1 multivariate normal vector $(Y_1, \ldots, Y_4)$ with $E(Y_t|X) = \beta_1 t + \beta_2 X$, $\beta_1 = 2$, $\beta_2 = 3$, $\text{var}(Y_t|X) = 5$ and $\text{cov}(Y_t, Y_{t'}|X) = 4$, $t \neq t'$. Then we generated R_4 given Y_4 iteratively for $t = 1, \ldots, 4$ from

$$\text{pr}(R_t = 0 | \bar{R}_{t-1}, \bar{L}_4) = \text{expit}[-2.25 + 0.25t + 0.5\{1 + I(t=1)\}X + \{2(1 - R_{t-1}) + 2R_{t-1}\varepsilon_{t-1}\} I(t>1) + \gamma\varepsilon_t],$$

where $\varepsilon_t = Y_t - \beta_1 t - \beta_2 X$ and $\gamma = 0$ or $\gamma = -0.5$. Under our data-generating process, model $\mathscr{B}(q)$ holds for $q_t(\bar{O}_{t-1}, Y_t) = \gamma Y_t$ and $h_t(\bar{O}_{t-1}; \alpha_t) = \alpha_{t0} + \alpha_{t1}X + I(t > 1)(\alpha_{t2}R_{t-1} + \alpha_{t3}R_{t-1}Y_{t-1})$ for a specific value of α_t. Since the functional form of $m_t(\bar{O}_{t-1})$ is complicated we have considered an approximate working model for it, given by $m_t(\bar{O}_{t-1}; \theta) = \theta_{t0} + \theta_{t1}X + I(t > 1)\{\theta_{t2}R_{t-1} + \theta_{t3}R_{t-1}X + \theta_{t4}R_{t-1}Y_{t-1}\}$, and thus computed a generalized 2^T-multiply robust estimator of (β_1, β_2) such that, at each t, model $\mathscr{C}_t(q)$ assumes that $m_t(\bar{O}_{t-1}) = m_t(\bar{O}_{t-1}; \theta)$ for some θ.

Under the data-generating process of the first experiment, the probability of not missing the outcome is 84.4, 69.0, 60.5 and 53.7 at the four occasions, respectively, for both $\gamma = 0$ and γ

= −0.5. In the second experiment, these probabilities are 84.4, 60.1, 61.5 and 56.5 for $\gamma = 0$. The values when $\gamma = -0.5$ are similar.

Table 1 summarizes the results for estimation of $\beta_3 = E(Y_3)$ and $\beta_4 = E(Y_4)$ in the first experiment for the following methods: inverse probability weighted estimators, i.e. those solving $E_n\{H_t(1, 0, \beta_t, \hat{\alpha}_t)\} = 0$, labelled 'IPW'; conditional mean imputation estimators solving $E_n\{M_t(1, 0, \beta_t, \hat{\theta}_t)\} = 0$, labelled 'CM'; and 2^T-multiply robust estimators, labelled 'MR'. The estimators were computed under various conditions: correctly specified working models $h_t(\bar{O}_{t-1}; \alpha_t)$ and $m_t(\bar{O}_{t-1}; \theta_t)$ as defined above for all t, labelled 'None'; models $m_3(\bar{O}_2; \theta_3)$ and $h_4(\bar{O}_3; \alpha_4)$ that incorrectly set a priori to zero the coefficients multiplying the term $R_{t-1}Y_{t-1}$, labelled '$\mathscr{C}_3$&$\mathscr{B}_4$'; and models $h_t(\bar{O}_{t-1}; \alpha_t)$ and $m_t(\bar{O}_{t-1}; \theta_t)$ that for all t incorrectly set a priori to zero the coefficients multiplying the term $R_{t-1}Y_{t-1}$, labelled 'All'.

The results for the 2^T-multiply robust estimators in the first simulation study are as predicted by the theory: they are nearly unbiased and the Wald confidence intervals centred at them cover roughly at the nominal 95% level when, at each occasion, none or one of the working models, but not both, is incorrect. In contrast, and also as expected, the inverse probability weighted estimators of β_3 are nearly unbiased but those of β_4 are not unbiased when the model for $h_3(\bar{O}_2)$ is correctly specified but that of $h_4(\bar{O}_3)$ is incorrectly specified. The reverse occurs for the conditional mean imputation estimators. No estimator is unbiased when all working models are misspecified. In addition, as predicted by theory, when $q_t = 0$, the 2^T-multiply robust estimator is more efficient than the inverse probability weighted estimator when all working models are correct, but is less efficient than the conditional mean imputation estimator. Interestingly, the 2^T-multiply robust estimators of β_3 and β_4 are also more efficient than the corresponding inverse probability weighted estimators when $q_t = -0.5Y_t$, and both working models are correct even though this cannot be deduced from the theory. Note also that the 2^T-multiply robust estimator of β_3 is less efficient than the inverse probability weighted estimator when $\mathscr{B}_3$ is correct but $\mathscr{C}_3$ is incorrect.

Table 2 summarizes the results from the second simulation for the generalized 2^T-multiply robust estimators of β_1 and β_2 solving the equations (18) that use, instead of $d(X)$, the vector function of X and β,

$$d^*(X;\beta) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ X & X & X & X \end{pmatrix} E_n\left\{Q^*\left(\beta, \tilde{\theta}, \tilde{\alpha}\right)^{\otimes 2}\right\},$$

where $Q^*(\beta, \theta, \alpha) = (Q_1(1, \beta, \theta_1, \alpha_1), \ldots, Q_T(1, \beta, \theta_T, \alpha_T))'$. It can be shown that using $d^*(X; \beta)$ in (18) results in generalized 2^T-multiply robust estimators that, under the union model $\cup_{u\in\mathbb{U}}\mathscr{M}_u^*(q)$, are asymptotically equivalent to those that solve (18) with $d(X) = d^*(X; \beta^*)$. The estimators were computed under various conditions: the correct working model $h_t(\bar{O}_{t-1}; \alpha_t)$ and the approximately correct model $m_t(\bar{O}_{t-1}; \theta)$, labelled 'None'; the same models as in the previous case except that the coefficients corresponding to $R_{t-1}Y_{t-1}$ of $h_2(\bar{O}_1; \alpha_2)$, $h_4(\bar{O}_3; \alpha_4)$ and $m_3(\bar{O}_2; \theta_3)$ were set to 0, labelled '$\mathscr{B}_2$, $\mathscr{C}_3$ and $\mathscr{B}_4$'; and the same models as in the first case but with the coefficients of $R_{t-1}Y_{t-1}$ equal to 0 for all t and for all models, labelled 'All'. The results confirm that the generalized 2^T-multiply robust estimators

perform well even if the model for $m_t(\bar{O}_{t-1})$ is incorrectly specified provided, as in our simulations, the model is richly parameterized.

Acknowledgement

This work was partially conducted while Stijn Vansteelandt was visiting the Department of Biostatistics of the Harvard School of Public Health. He wishes to thank the members of the Departments for their kind hospitality and stimulating environment. Stijn Vansteelandt was funded by a postdoctoral grant from the Fund for Scientific Research, Belgium. Andrea Rotnitzky and James Robins were funded by grants from the U.S. National Institutes of Health.

Appendix

Proofs

Proof of Theorem 1. Every observed data distribution is defined by a given collection of conditional densities and probabilities, $\{f(L_t|\bar{O}_{t-1}, R_t = 1), \text{pr}(R_t|\bar{O}_{t-1}) : t = 1, \ldots, T\}$. Thus, to show that $\mathscr{A}(q)$ is a model for the observed data law restricted only by $\text{pr}(R_t = 1|\bar{O}_{t-1}) > 0$ for $t = 1, \ldots, T$, it suffices to show the following: (a) given $\{f(L_t|\bar{O}_{t-1}, R_t = 1), \text{pr}(R_t|\bar{O}_{t-1}) : \text{pr}(R_t = 1|\bar{O}_{t-1}) > 0, t = 1, \ldots, T\}$, there exists a distribution $f^*(L_T, R_T)$ satisfying the restrictions of $\mathscr{A}(q)$ and such that, for $t = 1, \ldots, T$,

$$f(L_t|\bar{O}_{t-1}, R_t=1)=f^*(L_t|\bar{O}_{t-1}, R_t=1),\ \text{pr}(R_t=1|\bar{O}_{t-1})=\text{pr}^*(R_t=1|\bar{O}_{t-1}); \quad \text{(A1)}$$

and (b) $\text{pr}^*(R_t = 1|\bar{O}_{t-1}) > 0$ for every joint distribution $f^*(L_T, R_T)$ that satisfies the restrictions of model $\mathscr{A}(q)$.

We prove (a) by constructing a joint distribution $f^*(L_T, R_T)$ that satisfies (A1) iteratively as follows. For $t = 1$, R_{t-1} is nil and $f^*(L_{t-1}) = f(L_0)$. For each $t = 1, \ldots, T$, we define $f^*(L_t, R_t|L_{t-1}, R_{t-1})$ equal to $f^*(L_t, R_t|\bar{O}_{t-1})$, where $f^*(L_t, R_t|\bar{O}_{t-1})$ is defined by $f^*(L_t|\bar{O}_{t-1}, R_t = 1) = f(L_t|\bar{O}_{t-1}, R_t = 1)$ and $\text{pr}^*(R_t = 1|\bar{O}_{t-1}) = \text{pr}(R_t = 1|\bar{O}_{t-1})$, and when (4) holds $f^*(L_t \setminus Y_t|Y_t, \bar{O}_{t-1}, R_t = 0)$ is equal to an arbitrary law and $f^*(Y_t|\bar{O}_{t-1}, R_t = 0)$ is equal to the right-hand side of (5). By construction, $f^*(L_T, R_T)$ satisfies (1) and Assumption 1 for $t = 1, \ldots, T$, and thus is in model $\mathscr{A}(q)$, and additionally satisfies (A1), thus proving (a). Part (b) holds because it is implied by (1). This concludes the proof of part (i). To show part (ii), note that, under model $\mathscr{A}_t(q)$,

$$f\,(Y_t|\bar{O}_{t-1})=f(Y_t|\bar{O}_{t-1}, R_t=1)\ \left(\text{pr}(R_t=1|\bar{O}_{t-1})+\frac{\text{pr}(R_t=0|\bar{O}_{t-1})\ \exp\{q_t(\bar{O}_{t-1}, Y_t)\}}{E\,[\exp\{q_t(\bar{O}_{t-1}, Y_t)\}|\bar{O}_{t-1}, R_t=1]}\right),$$

where $q_t(\bar{O}_{t-1}, Y_t)$ is defined arbitrarily if $\text{pr}(R_t = 0|\bar{O}_{t-1}) = 0$. Thus, $f(Y_t|\bar{O}_{t-1})$ is identified under $\mathscr{A}_t(q)$ because the right-hand side of the last display is determined by the observed data law. To show that (8) holds, we use expression (7) and note that, under models $\mathscr{A}_t(q)$ and $\mathscr{A}(q)$, the right-hand side of (8) equals

$$E\left\{R_tY_t\left(1+\frac{\mathrm{pr}(R_t=0|\bar{O}_{t-1})\ \exp\{q_t(Y_t,\bar{O}_{t-1})\}}{\mathrm{pr}(R_t=1|\bar{O}_{t-1})E\left[\exp\{q_t(Y_t,\bar{O}_{t-1})\}|R_t=1,\bar{O}_{t-1}\right]}\right)\right\}$$
$$=E\left(\mathrm{pr}(R_t\right.$$
$$=1|\bar{O}_{t-1})E\left(Y_t|R_t\right.$$
$$\left.=1,\bar{O}_{t-1}\right)$$
$$+\mathrm{pr}(R_t$$
$$\left.=0|\bar{O}_{t-1})\frac{E\left[Y_t\ \exp\{q_t(Y_t,\bar{O}_{t-1})\}|R_t=1,\bar{O}_{t-1}\right]}{E\left[\exp\{q_t(Y_t,\bar{O}_{t-1})\}|R_t=1,\bar{O}_{t-1}\right]}\right).$$

From

$$\frac{E\left[Y_t\ \exp\{q_t(Y_t,\bar{O}_{t-1})\}|R_t=1,\bar{O}_{t-1}\right]}{E\left[\exp\{q_t(Y_t,\bar{O}_{t-1})\}|R_t=1,\bar{O}_{t-1}\right]}=E(Y_t|R_t=0,\bar{O}_{t-1})$$

under model $\mathscr{A}_t(q)$, it follows that the right-hand side is equal to $E(Y_t)$. The proof of (9) is now immediate.

Proof that the restrictions imposed by models $\mathscr{A}^*(q)$, $\mathscr{B}^*(q)$, $\mathscr{B}(q)$, $\mathscr{C}(q)$ *and* $\cap_{t=1}^{T}\{\mathscr{B}_t(q)\cap\mathscr{C}_t(q)\}$ *are compatible. To show that the restrictions of* $\mathscr{B}^*(q)$ *are compatible we will exhibit a joint law* $f^*(\bar{L}_T,\bar{R}_T)$ satisfying the restrictions defining $\mathscr{B}^*(q)$. We construct such a law recursively as follows. We define $f^*(L_0)$ as an arbitrary law and set R_0 as nil. Then, having defined $f^*(\bar{L}_{t-1},\bar{R}_{t-1})$ for $t=1,\ldots,T$, we define $f^*(L_t,R_t|\bar{L}_{t-1},\bar{R}_{t-1})$ as follows. The density $f^*(Y_t|\bar{L}_{t-1},\bar{R}_{t-1})$ satisfies $\bar{L}_{t-1}=\bar{L}_{t-1}\backslash X$ and the integral is taken with respect to the counting dominating measure for R_{t-1} and the adequate dominating measures for $\bar{L}_{t-1}$ and Y_t. This ensures that (17) holds. Next, we define $f^*(L_t\backslash Y_t|Y_t,R_t,\bar{O}_{t-1})$ as an arbitrary law. Finally, for a given fixed function $h_t(\bar{O}_{t-1};\alpha_t^*)$, we set

$$\mathrm{pr}^*(R_t=1|Y_t,\bar{L}_{t-1},\bar{R}_{t-1})=\mathrm{pr}^*(R_t=1|Y_t,\bar{O}_{t-1})=\left[1+\exp\{h_t(\bar{O}_{t-1};\alpha_t^*)+q_t(\bar{O}_{t-1},Y_t)\}\right]^{-1},$$

which ensures that (5) and (11) hold. Thus, by construction, $f^*(\bar{L}_T,\bar{R}_T)$ satisfies the restrictions defining $\mathscr{B}^*(q)$. Since $\mathscr{B}^*(q)$ is more restrictive than $\mathscr{B}(q)$ and $\mathscr{A}^*(q)$, this implies that the restrictions of $\mathscr{B}(q)$ and $\mathscr{A}^*(q)$ are compatible.

We next recursively construct a law $f^*(\bar{L}_T,\bar{R}_T)$ that satisfies the restrictions imposed by the intersection model $\cap_{t=1}^{T}\{\mathscr{B}_t(q)\cap\mathscr{C}_t(q)\}$. We define $f^*(L_0)$ as an arbitrary law and set R_0 as nil. Then, having defined $f^*(\bar{L}_{t-1},\bar{R}_{t-1})$ for $t=1,\ldots,T$, we define $f^*(L_t,R_t|\bar{L}_{t-1},\bar{R}_{t-1})$ as follows. Given fixed functions $h_t(\bar{O}_{t-1};\alpha_t^*)$ and $m_t(\bar{O}_{t-1};\theta_t^*)$, for each $t=1,\ldots,T$, we define $f^*(Y_t|\bar{L}_{t-1},\bar{R}_{t-1},R_t=0)=f^*(Y_t|R_t=0,\bar{O}_{t-1})$, where $f^*(Y_t|R_t=0,\bar{O}_{t-1})$ is any law that satisfies $\int Y_t f^*(Y_t|\bar{O}_{t-1},R_t=0)\mathrm{d}Y_t=m_t(\bar{O}_{t-1};\theta_t^*)$. Next, we define $\mathrm{pr}^*(R_t=1|\bar{L}_{t-1},\bar{R}_{t-1})=\mathrm{pr}^*(R_t=1|\bar{O}_{t-1})$ by the identity

$$\int \exp\{-h_t(\bar{O}_{t-1};\alpha_t^*)-q_t(\bar{O}_{t-1},Y_t)\}f^*(Y_t|\bar{O}_{t-1},R_t=0)\mathrm{d}Y_t=\frac{\mathrm{pr}^*(R_t=1|\bar{O}_{t-1})}{\mathrm{pr}^*(R_t=0|\bar{O}_{t-1})},$$

and we define $f^*(Y_t|\bar{L}_{t-1},\bar{R}_{t-1},R_t=1)=f^*(Y_t|\bar{O}_{t-1},R_t=1)$, where

$$f^*(Y_t|\bar{O}_{t-1},R_t=1)=\frac{\mathrm{pr}^*(R_t=0|\bar{O}_{t-1})}{\mathrm{pr}^*(R_t=1|\bar{O}_{t-1})}\frac{f^*(Y_t|\bar{O}_{t-1},R_t=0)}{\exp\{h_t|\bar{O}_{t-1};\alpha_t^*+q_t(\bar{O}_{t-1},Y_t)\}}.$$

Finally, we choose $f^*(L_t\setminus Y_t|Y_t,\bar{L}_{t-1},\bar{R}_{t-1})$ to be an arbitrary law. By construction, $f^*(\bar{L}_T,\bar{R}_T)$ satisfies for each t, (12), and hence model $\mathscr{C}_t(q)$, as well as (5) and (11), and hence model $\mathscr{B}_t(q)$. This is seen because, for the chosen law,

$$\exp\{h_t(\bar{O}_{t-1};\alpha_t^*+q_t(\bar{O}_{t-1},Y_t)\}=\frac{f^*(Y_t,R_t=0|\bar{O}_{t-1})}{f^*(Y_t,R_t=1|\bar{O}_{t-1})}=\frac{\mathrm{pr}^*(R_t=0|Y_t,\bar{O}_{t-1})}{\mathrm{pr}^*(R_t=1|Y_t,\bar{O}_{t-1})}.$$

We conclude that $f^*(\bar{L}_T,\bar{R}_T)$ satisfies the restrictions of model $\cap_{t=1}^T\{\mathscr{B}_t(q)\cap\mathscr{C}_t(q)\}$ and therefore also those of $\mathscr{C}(q)$.

Proof that $H_t(1,\varphi_{\theta_t\beta_t t},\beta_t,\alpha_t)=M_t(1,\psi_{\alpha_t,t},\beta_t,\theta_t)$. We have

$$\begin{aligned}
&\frac{R_t\varepsilon_t(\beta_t)}{\pi_t(\bar{O}_{t-1},Y_t;\alpha_t)}\\
&\quad+\left(1-\frac{R_t}{\pi_t(\bar{O}_{t-1},Y_t;\alpha_t)}\right)\{m_t(\bar{O}_{t-1};\theta_t)\\
&\quad-\beta_t\}=\left(1-\frac{R_t}{\pi_t(\bar{O}_{t-1},Y_t;\alpha_t)}\right)\{m_t(\bar{O}_{t-1};\theta_t)\\
&\quad-Y_t\}+Y_t-\beta_t=\left(1-R_t-\frac{R_t\{1-\pi_t(\bar{O}_{t-1},Y_t;\alpha_t)\}}{\pi_t(\bar{O}_{t-1},Y_t;\alpha_t)}\right)\{m_t(\bar{O}_{t-1};\theta_t)\\
&\quad-Y_t\}+R_t\varepsilon_t(\beta_t)+(1-R_t)(Y_t-\beta_t)=R_t\varepsilon_t(\beta_t)+(1-R_t)\{m_t(\bar{O}_{t-1};\theta_t)\\
&\quad-\beta_t\}+R_t\psi_{\alpha_t,t}(\bar{O}_{t-1})\exp\{q_t(\bar{O}_{t-1},Y_t)\}\{Y_t-m_t(\bar{O}_{t-1};\theta_t)\}.
\end{aligned}$$

Now, suppose that (11) and (4) hold. Then

$$E\left\{\frac{R_t\varepsilon_t(\beta_t^*)}{\pi_t(\bar{O}_{t-1},Y_t;\alpha_t^*)}\right\}=E\{\varepsilon_t(\beta_t^*)\}=0$$

and

$$E\left[\left\{1-\frac{R_t}{\pi_t(\bar{O}_{t-1},Y_t;\alpha_t^*)}\right\}\{m_t(\bar{O}_{t-1};\theta_t)-\beta_t^*\}\right]=E\left[E\left[\left\{1-\frac{R_t}{\pi_t(\bar{O}_{t-1},Y_t;\alpha_t^*)}\right\}\middle|\bar{O}_{t-1},Y_t\right]\{m_t(\bar{O}_{t-1};\theta_t^*)-\beta_t^*\}\right]=0$$

because

$$E\left\{\left(1-\frac{R_t}{\pi_t(\bar{O}_{t-1},Y_t;\alpha_t^*)}\right)|\bar{O}_{t-1},Y_t\right\}=0.$$

Thus, under (11) and (4), $E\{H_t(1,\phi_{\theta_t,\beta_t^*,t,\beta_t^*,\alpha_t^*})\}=0$ for any θ_t. Next, suppose (12) holds. Then $(1-R_t)\,m_t(\bar{O}_{t-1};\theta_t^*)=(1-R_t)\,E(Y_t|\bar{O}_{t-1},R_t=0)$, where $E(Y_t|\bar{O}_{t-1},R_t=0)$ is defined arbitrarily if (4) does not hold, and hence $E\left[R_t\varepsilon_t(\beta_t^*)+(1-R_t)\left\{m_t(\bar{O}_{t-1};\theta_t)-\beta_t^*\right\}\right]=E\{\varepsilon_t(\beta_t^*)\}=0$. Also,

$$E[R_t\psi_{\alpha_t,t}(\bar{O}_{t-1})\exp\{q_t(\bar{O}_{t-1},Y_t)\}\{Y_t-m_t(\bar{O}_{t-1};\theta_t^*)\}]=E[\psi_{\alpha_t,t}(\bar{O}_{t-1})E[\exp\{q_t(\bar{O}_{t-1},Y_t)\}\{Y_t-m_t(\bar{O}_{t-1};\theta_t^*)\}|\bar{O}_{t-1},R_t=1]\,\mathrm{pr}(R_t=1|\bar{O}_{t-1})]=0$$

because $E\left[\exp\{q_t(\bar{O}_{t-1},Y_t)\}\left\{Y_t-m_t(\bar{O}_{t-1};\theta_t^*)\right\}|\bar{O}_{t-1},R_t=1\right]=0$. Thus, under model (12), $E\{M_t(1,\psi_{\alpha_t,t},\beta_t^*,\theta_t^*)\}=0$ for any α_t.

Proof of Theorem 2. To prove part (i) of the theorem, we assume that the regularity conditions 1–9 of Appendix B of Robins et al. (1994) hold with $U_t(\beta_t,\theta_t,\alpha_t)$ and $(\beta_t^*,\theta_t^*,\alpha_t^*)$ replacing their $H_t(\gamma)$ and γ_0 respectively, and their regularity condition 3 being replaced by the assumption that $\mathrm{pr}(R_t=1|\bar{O}_{t-1},Y_t;\alpha_t)>\sigma>0$ with probabaility 1 for some σ and arbitrary α_t in the parameter space. By standard Taylor expansion arguments we have that $\sqrt{n}\left(\hat{\theta}_t-\theta_t^0\right)=-I_{\theta_t,M_t}^{-1}(\theta_t^0)\,n^{-1/2}\sum_i M_{it}(0,\psi_t,0,\theta_t^0)+o_p(1)$ and $\sqrt{n}(\hat{\alpha}_t-\alpha_t^0)=-I_{\alpha_t,H_t}^{-1}(\alpha_t^0)\,n^{-1/2}\sum_i H_{it}(0,\phi_t,0,\alpha_t^0)+o_p(1)$, where $o_p(1)$ denotes a random variable converging to 0 in probability. Furthermore, because $E\{Q_t(\beta_t^*,\theta_t^0,\alpha_t^0)\}=0$ under model $\mathscr{B}_t(q)\cup\mathscr{C}_t(q)$, another Taylor expansion gives

$$\begin{aligned}0=&n^{-1/2}\sum_i Q_{it}(\beta_t^*,\theta_t^0,\alpha_t^0)\\&+I_{\beta_t,Q_t}(\beta_t^*,\theta_t^0,\alpha_t^0)\sqrt{n}(\hat{\beta}_t-\beta_t^*)-I_{\alpha_t,Q_t}(\beta_t^*,\theta_t^0,\alpha_t^0)I_{\alpha_t,H_t}^{-1}(\alpha_t^0)\times n^{-1/2}\sum_i H_{it}(0,\phi_t,0,\alpha_t^0)\\&-I_{\theta_t,Q_t}(\beta_t^*,\theta_t^0,\alpha_t^0)I_{\theta_t,M_t}^{-1}(\theta_t^0)n^{-1/2}\sum_i M_{it}(0,\psi_t,0,\theta_t^0)+o_p(1).\end{aligned}$$

When, as in regularity condition 6 of Robins et al. (1994), $I_{\beta_t,Q_t}(\beta_t^*,\theta_t^*,\alpha_t^*)$ is nonsingular, this is equivalent to

$$\sqrt{n}(\hat{\beta}_t-\beta_t^*)=I_{\beta_t,Q_t}^{-1}(\beta_t^*,\theta_t^0,\alpha_t^0)n^{-1/2}\sum_i U_{it}(\beta_t^*,\theta_t^0,\alpha_t^0)+o_p(1).$$

The asymptotic distribution of $\sqrt{n}\left(\hat{\beta}_t - \beta_t^*\right)$ under model $\mathscr{B}_t(q) \cup \mathscr{C}_t(q)$ follows from the previous equation by Slutsky's Theorem and the Central Limit Theorem. The consistency of the variance estimator follows from the Law of Large Numbers. This proves part (i). Since α_t and θ_t, for $t = 1, \ldots, T$, are variation independent parameters, it also follows that $(I^{-1}_{\beta_1,Q_1}(\beta_1^*, \theta_1^0, \alpha_1^0)U_1(\beta_1^*, \theta_1^0, \alpha_1^0), \ldots, (I^{-1}_{\beta_T,Q_T}(\beta_T^*, \theta_T^0, \alpha_T^0)U_T(\beta_T^*, \theta_T^0, \alpha_T^0))'$ is the influence function corresponding to the 2^T-multiply robust CAN estimator for β* under model $\cap_{t=1}^{T}\{\mathscr{B}_t(q) \cup \mathscr{C}_t(q)\}$. This proves part (iii).

At the intersection model $\mathscr{B}_t(q) \cap \mathscr{C}_t(q)$, $I_{\alpha_t,Q_t}(\beta^*, \theta_t^0, \alpha_t^0) = I_{\theta_t,Q_t}(\beta^*, \theta_t^0, \alpha_t^0) = 0$ and hence $U_{\mathrm{it}}(\beta_t^*, \theta_t^0, \alpha_t^0) = Q_{\mathrm{it}}(\beta_t^*, \theta_t^0, \alpha_t^0)$. It follows that the estimators $\hat{\beta}_t\left(\psi_t^{(1)}, \phi_t^{(1)}\right)$ and $\hat{\beta}_t\left(\psi_t^{(2)}, \phi_t^{(2)}\right)$ have the same influence functions at the intersection model $\mathscr{B}_t(q) \cap \mathscr{C}_t(q)$. This proves part (ii). Part (iv) is similarly proved.

Proof of Theorem 3. By definition, $U(\beta, \theta^0, \alpha_t^0) = d(X)U^*(\beta, \theta_t^0, \alpha_t^0)'$ and, by analogous arguments to Theorem 2 for estimators $\hat{\beta}(d) \equiv \hat{\beta}(d, d_\theta, \psi, d_\alpha, \varphi)$, the variance matrix of the limiting distribution of $\sqrt{n}\{\hat{\beta}(d) - \beta^*\}$ is equal to $\Gamma(d) = \Psi(d)\,\Omega(d)\,\Psi(d)'$, where

$$\Psi(d) = E\left(d(X)\frac{\partial}{\partial \beta'}U^*(\beta, \theta_t^0, \alpha_t^0)'\Big|_{\beta=\beta^*}\right)^{-1}, \Omega(d) = E\left\{\left(d(X)U^*(\beta^*, \theta_t^0, \alpha_t^0)'\right)^{\otimes 2}\right\}$$

That $\Gamma(d_{\mathrm{opt}}) \leq \Gamma(d)$ follows after applying the Cauchy-Schwarz inequality.

References

Albert PS. A transitional model for longitudinal binary data subject to nonignorable missing data. Biometrics. 2000; 56:602–608. [PubMed: 10877323]

Andersson SA, Perlman MD. Lattice-ordered conditional-independence models for missing data. Statist. Prob. Lett. 1991; 12:465–486.

Deltour I, Richardson S, Le Hesran J-Y. Stochastic algorithms for Markov models estimation with intermittent missing data. Biometrics. 1999; 55:565–573. [PubMed: 11318215]

Fairclough DL, Peterson HF, Cella D, Bonomi P. Comparison of several model-based methods for analysing incomplete quality of life data in cancer clinical trials. Statist. Med. 1998; 17:781–796.

Gill, RD.; Robins, JM. Sequential models for coarsening and missingness. In: Lin, DY.; Fleming, TR., editors. Proc. First Seattle Symp. Biostatist: Survival Anal. New York: Springer; 1997. p. 295-305.

Gill, RD.; van der Laan, MJ.; Robins, JM. Coarsening at random: characterizations, conjectures and counterexamples. In: Lin, DY.; Fleming, TR., editors. Proc. First Seattle Symp. Biostatist: Survival Anal. New York: Springer; 1997. p. 255-294.

Ibrahim JG, Chen M-H, Lipsitz SR. Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. Biometrika. 2001; 88:551–564.

Laird N, Ware J. Random effects models for longitudinal data. Biometrics. 1982; 38:963–974. [PubMed: 7168798]

Lin H, Scharfstein DO, Rosenheck RA. Analysis of longitudinal data with irregular, informative follow-up. J. R. Statist. Soc. B. 2003; 66:791–813.

Little, RJ.; Rubin, DB. Statistical Analysis with Missing Data. New York: Wiley; 1987.

Robins JM. Non-response models for the analysis of non-monotone nonignorable missing data. Statist. Med. 1997; 16:21–37.

Robins, JM. Proc. Am. Statist. Assoc. Sec. Bayesian Sci. Alexandria: Am. Statist. Assoc; 2000. Robust estimation in sequentially ignorable missing data and causal inference models; p. 6-10.1999

Robins JM, Gill RD. Non-response models for the analysis of non-monotone ignorable missing data. Statist. Med. 1997; 16:39–56.

Robins, JM.; Rotnitzky, A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell, N.; Dietz, K.; Farewell, V., editors. AIDS Epidemiology - Methodological Issues. Boston, MA: Birkhäuser; 1992. p. 297-331.

Robins JM, Rotnitzky A. Bickel P, Kwon J. Statist. Sinica. 2001; 11:920–936. Comment on a paper by.

Robins, JM.; Rotnitzky, A.; Scharfstein, D. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran, ME.; Berry, D., editors. Statistical Models in Epidemiology: The Environment and Clinical Trials. Vol. IMA Volume 116. New York: Springer-Verlag; 1999. p. 1-92.

Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J. Am. Statist. Assoc. 1994; 89:846–866.

Robins JM, Rotnitzky A, Zhao L-P. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J. Am. Statist. Assoc. 1995; 90:106–121.

Rotnitzky A, Robins JM, Scharfstein DO. Semiparametric regression for repeated outcomes with nonignorable nonresponse. J. Am. Statist. Assoc. 1998; 93:1321–1339.

Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric non-response models. J. Am. Statist. Assoc. 1999; 94:1096–1146.

Shah A, Laird N, Schoenfeld D. A random-effects model for multiple characteristics with possibly missing data. J. Am. Statist. Assoc. 1997; 92:775–779.

Troxel AB, Fairclough DL, Curran D, Hahn EA. Statistical analysis of quality of life with missing data in cancer clinical trials. Statist. Med. 1998; 17:653–666.

Troxel AB, Lipsitz SR, Harrington DP. Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. Biometrika. 1998; 85:661–672.

van der Laan, MJ.; Robins, JM. Unified Methods for Censored Longitudinal Data and Causality. New-York: Springer-Verlag; 2003.

Zeuzem S, Feinman SV, Rasenack J, Heathcote EJ, Lai MY, Gane E, O'Grady J, Reichen J, Diago M, Lin A, Hoffman J, Brunda MJ. Peginterferon alfa-2a in patients with chronic hepatitis C. New Engl. J. Med. 2000; 343:1666–1672. [PubMed: 11106715]

Table 1

Results from the first simulation study. Bias is 100 times the Monte Carlo mean minus true parameter value, SD is 100 times the Monte Carlo standard deviation, Cov is the Monte Carlo coverage probability times 100.

			IPW			CM			MR		
q_t	Estimand	Misspec.	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov
0	$\beta_3 = 6.69$	None	−1.6	27.7	92.6	0.1	16.4	95.1	−0.1	16.8	94.3
		C_3 & B_4	−1.6	27.7	92.6	−48.8	16.2	14.4	−0.6	30.7	93.8
		All	−52.2	17.7	15.8	−48.8	16.4	14.4	−48.9	16.2	15.4
	$\beta_4 = 8.61$	None	−1.3	28.5	91.5	0.2	16.4	95.4	0.3	16.8	95.5
		C_3 & B_4	−52.1	18.4	18.7	0.2	16.4	95.4	0.3	16.4	95.4
		All	−52.1	18.4	18.7	−47.6	16.4	18.3	−46.7	16.2	19.4
$-0.5\,Y_t$	$\beta_3 = 6.99$	None	−0.4	28.5	93.0	−2.3	22.0	89.5	1.0	17.9	95.0
		C_3 & B_4	−0..4	28.5	93.0	−86.2	29.5	11.6	0.7	44.1	92.9
		All	−82.1	19.4	1.0	−86.2	29.5	11.6	−77.3	17.9	0.9
	$\beta_4 = 8.87$	None	−0.7	27.6	92.2	−4.0	23.9	94.1	1.2	17.3	94.9
		C_3 & B_4	−81.3	20.2	2.2	−4.0	23.9	94.1	−1.0	18.8	94.1
		All	−81.3	20.2	2.2	−86.9	34.7	15.2	−74.6	18.3	1.4

IPW, inverse probability weighted estimator; CM, conditional mean imputation estimator; MR, 2^T-multiply robust estimator; None, no model misspecification; C_3 & B_4, misspecification of models C_3 and B_4; All, misspecification of C_t and B_t for each t.

Table 2

Results from the second simulation study. Bias is 100 times the Monte Carlo mean minus true parameter value, SD is 100 times the Monte Carlo standard deviation, Cov is the Monte Carlo coverage probability times 100.

q_t	Estimand	Misspec.	Bias	SD	Cov
0	$\beta_1 = 3$	None	0.08	10.60	94.6
		B_2, C_3 & B_4	0.12	10.30	94.0
		All	1.74	9.58	95.7
	$\beta_2 = 2$	None	0.007	4.25	95.3
		B_2, C_3 & B_4	−0.66	3.68	95.9
		All	−10.7	3.87	17.2
$-0.5\,Y_t$	$\beta_1 = 3$	None	1.70	12.00	95.7
		B_2, C_3 & B_4	1.36	12.20	93.1
		All	3.37	11.30	91.2
	$\beta_2 = 2$	None	1.31	6.43	96.7
		B_2, C_3 & B_4	−4.21	7.69	88.1
		All	−24.8	9.10	5.71

None, no model misspecification; B_2, C_3 & B_4, misspecification of models B_2, C_3 and B_4; All, misspecification of C_t and B_t for each t.