



Invited Commentary

Invited Commentary: Effect Modification by Time-varying Covariates

James M. Robins^{1,2}, Miguel A. Hernán¹, and Andrea Rotnitzky^{2,3}

¹ Department of Epidemiology, Harvard School of Public Health, Boston, MA.

² Department of Biostatistics, Harvard School of Public Health, Boston, MA.

³ Department of Economics, Universidad Di Tella, Buenos Aires, Argentina.

Received for publication November 21, 2006; accepted for publication March 9, 2007.

Marginal structural models (MSMs) allow estimation of effect modification by baseline covariates, but they are less useful for estimating effect modification by evolving time-varying covariates. Rather, structural nested models (SNMs) were specifically designed to estimate effect modification by time-varying covariates. In their paper, Petersen et al. (*Am J Epidemiol* 2007;000:000–00) describe history-adjusted MSMs as a generalized form of MSM and argue that history-adjusted MSMs allow a researcher to easily estimate effect modification by time-varying covariates. However, history-adjusted MSMs can result in logically incompatible parameter estimates and hence in contradictory substantive conclusions. Here the authors propose a more restrictive definition of history-adjusted MSMs than the one provided by Petersen et al. and compare the advantages and disadvantages of using history-adjusted MSMs, as opposed to SNMs, to examine effect modification by time-dependent covariates.

causality; confounding factors (epidemiology); longitudinal studies; nested model; observational data; structural model; time-dependent covariate

Abbreviations: MSM, marginal structural model; SNM, structural nested model.

Marginal structural models (MSMs) are being increasingly used to estimate the effects of time-varying treatments or exposures. Unlike conventional statistical methods, MSMs allow consistent estimation of the effect of a time-varying treatment on an outcome of interest even when there is confounding by time-varying covariates affected by earlier treatment. However, MSMs have an important limitation. As was pointed out by Robins (1, 2) and Hernán et al. (3), MSMs naturally allow estimation of effect modification by baseline covariates, but they are less useful for estimating effect modification by evolving time-varying covariates. Rather, structural nested models (SNMs) were specifically designed to estimate effect modification by time-varying covariates.

In this issue of the *Journal*, Petersen et al. (4) describe history-adjusted MSMs, a generalized form of MSM that was first proposed by Joffe et al. (5) and studied in detail by van der Laan et al. (6). Petersen et al. argue that history-

adjusted MSMs allow a researcher to easily estimate effect modification by time-varying covariates, thus overcoming an important shortcoming of standard MSMs.

However, as we explain below, this apparent advantage of history-adjusted MSMs over standard MSMs comes at a price: History-adjusted MSMs can produce logically incompatible parameter estimates and hence result in contradictory substantive conclusions. As a consequence, clinicians or other decision-makers relying on history-adjusted MSMs to decide the best course of action can be left without guidance. In this commentary, we clarify how history-adjusted MSMs differ from standard MSMs and describe the conditions under which incompatible parameter estimates can arise in the former. We also propose a more restrictive definition of history-adjusted MSMs than the one provided by Petersen et al. (4) and compare the advantages and/or disadvantages of using history-adjusted MSMs, as opposed to

SNMs, to examine effect modification by time-dependent covariates.

STANDARD MARGINAL STRUCTURAL MODELS

We start by briefly reviewing standard MSMs using Petersen et al.'s notation. To simplify the exposition, we assume a closed cohort with a well-defined time of enrollment for each subject and no loss to follow-up. Time is measured in periods (e.g., months) since time of enrollment, $m = 0$, until the end of follow-up, $m = K + 1$. We denote the treatment received in month m as $A(m)$ and covariates measured at the start of month m as $L(m)$. A subject's chronologically ordered data are therefore $L(0), A(0), L(1), A(1), \dots, L(K), A(K), L(K + 1)$. In Petersen et al.'s article (4), a subject's $A(m)$ is 1 for the times m that the subject stays on the failing antiretroviral treatment and 0 after switching to another treatment, and CD4 T-cell count $Y(m)$ is a component of the vector $L(m)$. A nondynamic treatment regime that specifies the treatment at each time from time m through time $t - 1$ is denoted by $\underline{a}(m, t - 1) = \{a(m), a(m + 1), \dots, a(t - 1)\}$. For example, in the paper by Petersen et al. (4), $\underline{a}(m, t - 1) = \{1, 1, 0, 0, 0, \dots, 0\}$ would be the regime "switch from the failing treatment at time $m + 2$ and continue on the new treatment through $t - 1$." The counterfactual (or potential) variable $Y_{\underline{a}(m)}(t)$ represents a subject's CD4 T-cell count measured at time t had the subject followed regime $\underline{a}(m, t - 1)$. In the paper by Petersen et al. (4), t is $m + 8$.

A standard MSM can be used to model the mean CD4 T-cell count $Y_{\underline{a}(m)}(t)$ at time t under all possible nondynamic treatment regimes from baseline time m to $t - 1$, that is, $E[Y_{\underline{a}(m)}(t)]$, where $E[X]$ is the expected value or mean of the random variable X . If so desired, the model may be made conditional on baseline variables $V(m)$ to model the conditional mean $E[Y_{\underline{a}(m)}(t)|V(m)]$ as a function of $\underline{a}(m, t - 1)$ and $V(m)$. Here $V(m)$ is a vector whose components may include any function of a subject's treatment and covariate history measured before $A(m)$. The model is not defined until the analyst chooses a baseline time m , a response time t , and a functional form for $E[Y_{\underline{a}(m)}(t)|V(m)]$. The choice of the times m and t turns out to be a key point in the comparison of standard versus history-adjusted MSMs, so we defer the discussion of this topic to the next section. For now, let us think of m and t as two fixed times after the time of enrollment—for example, $m = 1$ and $t = 9$. As to the choice of a functional form for $E[Y_{\underline{a}(m)}(t)|V(m)]$, the analyst needs to use her subject-matter knowledge to decide what functions of treatment (e.g., duration of treatment, average treatment dose) and baseline variables $V(m)$ are the most appropriate.

An example of a standard MSM is

$$\begin{aligned} E[Y_{\underline{a}(m)}(t)|V(m)] \\ &= \theta_0 + \theta_1 V(m) + \beta_0 \text{dur}[\underline{a}(m, t - 1)] \\ &\quad + \beta_1 \text{dur}[\underline{a}(m, t - 1)]V(m), \end{aligned}$$

where

$$\text{dur}[\underline{a}(m, t - 1)] = \sum_{j=m}^{t-1} a(j)$$

is the duration of use of the failing treatment from the baseline time m through $t - 1$, as described in the paper by Petersen et al. (4). The parameter vector $\beta = (\beta_0, \beta_1) = (0, 0)$ is equivalent to the null hypothesis that treatment has no effect—that is, that $E[Y_{\underline{a}(m)}(t)|V(m)]$ is the same for all regimes $\underline{a}(m, t - 1)$. The parameter β_1 captures effect modification by baseline variables on an additive scale: If β_0 and $\beta_0 + \beta_1 v(m)$ differ in sign for certain values $v(m)$ of $V(m)$, there is qualitative effect modification by $V(m)$. In particular, if $V(m)$ is univariate and binary, there is qualitative effect modification by $V(m)$ if β_0 and $\beta_0 + \beta_1$ differ in sign. Under the assumption of no unmeasured confounding for the effect of treatment from time m to $t - 1$ on the mean of $Y(t)$, the parameters of the MSM can be consistently estimated by inverse probability weighting (see Appendix).

Before comparing standard and history-adjusted MSMs in the next section, we point out one important warning for the causal interpretation of a standard MSM. Suppose the baseline time m exceeds 0 and $V(m)$ is a vector with two components: "duration of treatment before m " and "CD4 T-cell count at m ." Further suppose that both the estimate of the main effect of "treatment duration before m " and the estimate of the interaction between "treatment duration before m " and $\text{dur}[\underline{a}(m, t - 1)]$ ("treatment duration from m onwards") are large and highly significant. One cannot conclude that "treatment duration before m " has a causal effect on the response $Y(t)$, because these results are compatible with 1) unmeasured confounding for treatment before m or 2) selection bias. To understand why those results might be explained by selection bias, consider the following scenario: 1) Treatment before m is a cause of CD4 T-cell count at m but not a cause of CD4 T-cell count at t , $Y(t)$, and 2) an unmeasured genetic trait that is unassociated with treatment history is a cause of CD4 T-cell count at m and also causes $Y(t)$ both directly and by interacting with treatment subsequent to m . When conditions 1 and 2 hold, conditioning the analysis on CD4 T-cell count at m , a common effect of the genetic trait and treatment before m , induces an association between treatment before m and the unmeasured genetic trait, and therefore between treatment before m and $Y(t)$ (7). The causal directed acyclic graph shown in figure 1 depicts this situation with $A(m-)$, $C(m)$, $A(m+)$, and $Y(t)$ representing treatment before baseline m , CD4 T-cell count at m , treatment after baseline m , and outcome at time t , respectively. We refer to this association as "selection bias" because it exists even when both treatment before m has no causal effect on $Y(t)$ and the genetic trait responsible for the selection bias is marginally unassociated with treatment before m and thus is a nonconfounder.

STANDARD VERSUS HISTORY-ADJUSTED MARGINAL STRUCTURAL MODELS

Below we discuss the choice of the response time t and the baseline time m . As we will see, these choices are intimately connected with the definitions of standard and history-adjusted MSMs.

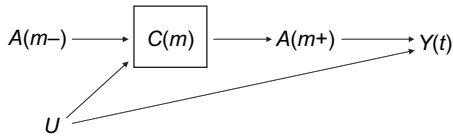


FIGURE 1. Selection bias for the effect of treatment before baseline.

Let us first discuss the choice of the response time t . The above standard MSM models the mean outcome at a single fixed time t (e.g., $t = 9$), and thus we say that it is a univariate MSM. However, a standard MSM need not be univariate. If we are willing to assume that the above model holds for all possible values of t greater than baseline time m , we can simultaneously model the mean of the outcome at all times $t > m$. The MSM is then multivariate. The procedure for the estimation, via inverse probability weighting, of the parameters of a multivariate MSM requires only a minor generalization of the procedure used for univariate MSMs (see Hernán et al. (8) and the Appendix for details). If one believes this multivariate MSM to be unrealistic because it assumes that the effect of treatment does not depend on the time t , one can make the model more flexible and allow for treatment effects that vary with time by replacing $\theta = (\theta_0, \theta_1)$ and $\beta = (\beta_0, \beta_1)$ with time-specific parameter vectors, $\theta_t = (\theta_{0,t}, \theta_{1,t})$ and $\beta_t = (\beta_{0,t}, \beta_{1,t})$.

Let us now turn our attention to the choice of the baseline time m . In many longitudinal studies, the effect of treatment received at time m from enrollment will be confounded unless one can adjust for high-quality time-varying laboratory, clinical, and treatment data collected over a number of periods prior to m . Any time m at which such high-quality data are available is eligible to be the baseline time of an MSM, although generally the earliest eligible time is chosen. For example, if measurements of treatment in the past month and CD4 T-cell counts in the previous 2 months were needed to control confounding for the effect of current treatment, then the earliest possible baseline time would be $m = 1$ if CD4 T-cell measurements began at the time of enrollment $m = 0$. (See Robins et al. (9) for a more detailed discussion.)

However, rather than using precisely one eligible baseline time (e.g., $m = 1$), one could decide to use all eligible baseline times $m = 1, 2, 3, \dots$ before t . Thus, in the above univariate MSM, we could see m as an index for multiple baseline times instead of as a single fixed time m . If one believes the MSM with multiple baseline times to be unrealistic because it assumes that the effect of treatment does not depend on the baseline time m , one can make the model more flexible and allow for treatment and covariate effects that vary with time by replacing $\theta = (\theta_0, \theta_1)$ and $\beta = (\beta_0, \beta_1)$ with time-specific parameter vectors, $\theta_m = (\theta_{0,m}, \theta_{1,m})$ and $\beta_m = (\beta_{0,m}, \beta_{1,m})$, which are indexed by the eligible baseline times $m < t$.

A univariate MSM with multiple baseline times appears closely analogous to a multivariate MSM, except with multiple baseline times m per subject substituted for multiple response times t . In fact, the procedure for estimation, via inverse probability weighting, of the parameters of a univariate

MSM with multiple baseline times and of a multivariate MSM are also analogous (see Appendix).

Petersen et al. (4) refer to MSMs with multiple baseline times as “history-adjusted MSMs” (6, 10). MSMs with multiple baseline times can be divided into two mutually exclusive groups. For a given MSM and outcome time t , let $\text{num}(t)$ count the number of different baseline times m for which the MSM models the effect of regimes beginning at m on the outcome $Y(t)$. The first group is composed of MSMs for which $\text{num}(t)$ exceeds 1 for one or more outcome times t . This group includes MSMs, similar to those considered in an earlier paper by van der Laan et al. (6), that model the effect on an outcome $Y(t)$ of treatment regimes beginning at all times m prior to t . The second group is composed of MSMs for which $\text{num}(t)$ is 1 for all outcome times t . This group includes the MSM discussed by Petersen et al. (4) that restricts the set of outcome times to months 8 and later and only models the effect on each outcome $Y(t)$ of treatment regimes beginning at time $m = t - 8$. We propose that the use of the term “history-adjusted MSM” be reserved for the first group, for the following reasons.

First, restricting the name “history-adjusted” to MSMs in group 1 is more in keeping with Petersen et al.’s conceptualization of the difference between history-adjusted MSMs and standard MSMs (4, 10). Specifically, in their abstract, the authors state that “unlike standard MSMs, history-adjusted MSMs can be used to estimate modification of treatment effects by time-varying covariates” (4, p. 000). However, this claim is true only for MSMs in group 1: To estimate effect modification by a time-varying covariate on a response $Y(t)$, we must, by definition, model effect modification at two or more times m , since otherwise we could regard the covariate as non-time-varying. In contrast to MSMs in group 1, MSMs in group 2 are like standard MSMs in that, for a given response $Y(t)$, they estimate the magnitude of effect modification by past time-varying covariates only at a single baseline time m —for example, $m = t - 8$. For this reason, we can regard MSMs in group 2 to be simply a collection of ordinary MSMs that, just like multivariate MSMs, allow increased estimation efficiency 1) by assuming that the parameters corresponding to different members of the collection are related and 2) by using more realistic working models than the independence model for within-individual correlations.

Second, it was only MSMs in group 1 that Robins (1, 2) was warning against when he stated that MSMs could not be easily used to estimate effect modification by evolving time-dependent covariates. This is because, as we explain below and in the Appendix, only MSMs in group 1 can be incompatible and thus lead to logical inconsistencies. Henceforth we refer only to models in group 1 as history-adjusted MSMs.

MODEL INCOMPATIBILITY IN HISTORY-ADJUSTED MARGINAL STRUCTURAL MODELS

Below we show that the apparently nearly exact analogy between history-adjusted MSMs and a standard multivariate MSM goes only so far. Specifically, a history-adjusted MSM, unlike a standard multivariate MSM, may be an

incompatible model. We say a model is incompatible if there exist any logically inconsistent (incompatible) parameter values. A familiar case of an incompatible model is a linear regression model $\Pr(D = 1|X) = \alpha_0 + \alpha_1 X$ for a binary outcome D . For example, if the covariate X takes values $0, 1, \dots, 100$ and one fits this model by a method (such as ordinary least squares) that does not impose the constraint that predicted probabilities must lie between 0 and 1, one can easily obtain incompatible parameter estimates, such as $\hat{\alpha}_0 = 0, \hat{\alpha}_1 = 0.02$, that result in illogical statements such as “ $2 = 0 + (0.02)100$ is the estimated probability that $D = 1$ among subjects with $X = 100$.” In contrast, a logistic regression model $\text{logit } \Pr(D = 1|X) = \alpha_0 + \alpha_1 X$ is compatible, because $e^{\alpha_0 + \alpha_1 X} / (1 + e^{\alpha_0 + \alpha_1 X})$ is always between 0 and 1.

We now provide an informal explanation of why history-adjusted MSMs may be incompatible (see the Appendix for a formal treatment). Let us start by considering our original univariate MSM with the only response time t equal to $K + 1$ but now with multiple baseline times m :

$$\begin{aligned} E[Y_{\underline{a}(m)}(K+1)|V(m)] \\ = \theta_0 + \theta_1 V(m) + \beta_0 \text{dur}[\underline{a}(m, K)] \\ + \beta_1 \text{dur}[\underline{a}(m, K)]V(m), \end{aligned}$$

where $V(m)$ is the entire covariate and treatment history measured before $A(m)$. This history-adjusted MSM makes three critical assumptions:

1. The direct effect of baseline treatment $a(m)$ is the same as the effect of each subsequent component of the treatment.
2. The effect of treatment from $m + 1$ to K is the same regardless of (i.e., is not modified by) the value of the baseline treatment $a(m)$.
3. The effect of (baseline and subsequent) treatment is the same for all baseline times.

In many settings, including most human immunodeficiency virus studies like the one described by Petersen et al. (4), these three assumptions are implausible and a more flexible, realistic, history-adjusted MSM is needed, such as

$$\begin{aligned} E[Y_{\underline{a}(m)}(K+1)|V(m)] \\ = \theta_0 + \theta_1 V(m) + \beta_0^{(1)} a(m) + \beta_1^{(1)} a(m)V(m) \\ + \beta_2^{(1)} a(m)(K-m) + \beta_0^{(2)} \text{dur}[\underline{a}(m+1, K)] \\ + \beta_1^{(2)} \text{dur}[\underline{a}(m+1, K)]V(m) + \beta_2^{(2)} \text{dur}[\underline{a}(m+1, K)] \\ \times (K-m) + \beta_3^{(2)} \text{dur}[\underline{a}(m+1, K)]a(m). \end{aligned}$$

This model relaxes assumption 1 by including separate parameters for $a(m)$ and $\text{dur}[\underline{a}(m+1, K)]$, assumption 2 by including an interaction term between $a(m)$ and $\text{dur}[\underline{a}(m+1, K)]$, and assumption 3 by including an interaction term between time to the end of follow-up $K - m$ and both $a(m)$ and $\text{dur}[\underline{a}(m+1, K)]$. In this model, the parameter vector $\beta^{(1)} = (\beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)})$ encodes the direct effect of baseline treatment when subsequent treatment is withheld, $\underline{a}(m+1, K) = 0$. We will henceforth refer to this simply as the direct effect of baseline treatment. The parameter vector $\beta^{(2)} =$

$(\beta_0^{(2)}, \beta_1^{(2)}, \beta_2^{(2)}, \beta_3^{(2)})$ encodes the effect of subsequent cumulative treatment. The estimates of the model parameters, $(\hat{\theta}, \hat{\beta}^{(1)}, \hat{\beta}^{(2)})$, can be obtained by inverse probability weighting. The problem with this more flexible, realistic model is that it may lead to parameter estimates that are logically inconsistent, as we discuss below.

Suppose, as an example, that 1) the components of $\hat{\beta}^{(1)}$ are all negative and lie within the interval $(-1/4, -3/4)$ and a joint 95 percent confidence interval for $\beta^{(1)}$ only includes vectors with all components lying between -1 and -0.01 and 2) the components of $\hat{\beta}^{(2)}$ all exceed 10 and a joint 95 percent confidence interval for $\beta^{(2)}$ includes only vectors with all components exceeding 8. The negative $\hat{\beta}^{(1)}$ implies that, for each m , the effect of baseline treatment $a(m)$ has a negative effect on $Y(K+1)$ when $\underline{a}(m+1, K) = 0$. The positive $\hat{\beta}^{(2)}$ implies that cumulative treatment from $m+1$ to K has a large positive effect. Furthermore, suppose the confidence intervals imply that the opposite signs of the estimated effects of baseline versus subsequent treatment cannot be explained by sampling variability.

However, it is logically impossible for cumulative treatment from $m+1$ to K to have a large positive effect on $Y(K+1)$ if, for each time s greater than m , $a(s)$ alone has a negative effect. This implies that the history-adjusted MSM is an incompatible model and the parameter estimates $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ are logically inconsistent. This result is made precise in theorem 1, shown in the Appendix, where it is formally proven that pairs $(\beta^{(1)}, \beta^{(2)})$ with $\beta^{(1)}$ negative and $\beta^{(2)}$ positive are logically incompatible.

The incompatible estimates of $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ also result in logically inconsistent statements about clinical strategies. Specifically, theorem 1 shows that all components of $\hat{\beta}^{(1)}$ being less than 0 implies that the estimated optimal treatment regime starting from any eligible time m is the regime $\underline{0}(m)$, “always withhold treatment from m .” However, in the Appendix, we also show that all components of $\hat{\beta}^{(2)}$ being positive and larger in absolute value than those of $\hat{\beta}^{(1)}$ implies that the regime $\underline{1}(m)$, “always take treatment starting at m ,” is (estimated) to be preferable to the regime $\underline{0}(m)$. The preceding two statements are logically inconsistent and taken together would leave a health-care provider without any guidance as to a reasonable treatment strategy. Thus, an analyst committed to using history-adjusted MSMs would face two undesirable alternatives: to use a compatible but unrealistic, and therefore probably very badly misspecified, model or to use a more realistic but incompatible model that may lead to logically inconsistent estimates.

Of course, the use of incompatible models only poses a difficulty if incompatible estimates are likely to occur. It is clear that an ordinary least-squares fit of our linear Bernoulli regression model will frequently result in incompatible estimates. It may be less clear that an inverse probability weighting fit of our incompatible history-adjusted MSM can also easily result in incompatible estimates. However, in the model used in our example, incompatible estimates may occur if 1) the model is somewhat misspecified in that the effect of subsequent treatment on the mean outcome actually depends on a much more complicated function of $\underline{a}(m+1, K)$ than the assumed linear dependence on $\text{dur}[\underline{a}(m+1, K)]$ and 2) for most times j , $A(j)$ is highly

correlated with the part of that complicated function of $A(m+1, K)$ that is uncorrelated with $\text{dur}[A(m+1, K)]$. In the Appendix, we argue that it may be prohibitively difficult to develop an empirical test of fit for a history-adjusted MSM that reliably indicates that conditions 1 and 2 have not only occurred but are of sufficient magnitude to produce estimates which suffer from incompatibility to such an extent that the clinically relevant inferences may be compromised.

STRUCTURAL NESTED MODELS VERSUS HISTORY-ADJUSTED MARGINAL STRUCTURAL MODELS

We have seen that the problem in fitting a history-adjusted MSM by inverse probability weighting is that the estimates $\hat{\beta}^{(1)}$ of the direct effect of baseline treatment $a(m)$ may be logically inconsistent with the estimated effect $\hat{\beta}^{(2)}$ of subsequent treatment. A natural way to overcome this difficulty would be to only model the direct effect of $a(m)$ while leaving the effect of subsequent treatment unmodeled. When, as in our example, $V(m)$ is the entire covariate-and-treatment history measured before $A(m)$, a model for the direct effect of $a(m)$ at all eligible baseline times m is precisely an (additive) SNM (11, 12). Because we are only modeling the direct effect of baseline treatment, the parameters $\beta^{(1)}$ can no longer be well estimated by inverse probability weighting but rather should be estimated using g -estimation. Furthermore, after obtaining a g -estimate $\hat{\beta}^{(1)}$, one can estimate the mean of the counterfactual outcome of interest under any treatment regime without having to model the effect of subsequent treatment and thus without having to use incompatible models (see Appendix).

Indeed, the fact that we can estimate $\beta^{(1)}$ by g -estimation rather than inverse probability weighting is a second important benefit (in addition to avoiding model incompatibility) of using an SNM rather than a history-adjusted MSM. As we describe in the Appendix, inverse probability weighting estimation requires a “positivity assumption” (13) and is sensitive to the presence of extreme weights, either true or estimated. In contrast, g -estimation does not require a positivity assumption and is much less affected by extreme weights. The Appendix also contains a brief discussion of approaches other than g -estimation to handling model incompatibility and of how incompatible models might be used for goodness-of-fit testing and model selection.

In the absence of model misspecification or confounding by unmeasured factors, both inverse probability weighting estimation of standard or history-adjusted MSMs and g -estimation of SNMs allow one to estimate the effect of a time-varying treatment even when there is time-dependent confounding by time-varying covariates affected by earlier treatment. However, for the reasons discussed above, we would recommend that SNMs rather than history-adjusted MSMs be the routine model choice for investigation of effect modification by evolving time-varying covariates. Nevertheless, we also encourage comparison of the results obtained with SNMs to those obtained with history-adjusted MSMs to deepen our understanding of and experience with these new models. Only through such comparisons will we learn whether incompatible estimates occur with history-adjusted MSMs frequently enough to be of concern.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants R37-AI032475 and R01-HL080644.

Conflict of interest: none declared.

REFERENCES

1. Robins JM. Marginal structural models. In: 1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science. Alexandria, VA: American Statistical Association, 1998:1–10.
2. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran E, Berry D, eds. *Statistical models in epidemiology: the environment and clinical trials*. New York, NY: Springer-Verlag, 1999:95–134.
3. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of human immunodeficiency virus-positive men. *Epidemiology* 2000;11:561–70.
4. Petersen M, Deeks S, Martin J, et al. History-adjusted marginal structural models for estimating time-varying effect modification. *Am J Epidemiol* 2007;000:000–00.
5. Joffe M, Santanna J, Feldman H. Partially marginal structural models for causal inference. (Abstract). *Am J Epidemiol* 2001;153(suppl):S261.
6. van der Laan MJ, Petersen ML, Joffe MM. History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *Int J Biostat* 2005;1:article 4. (Electronic article). (<http://www.bepress.com/ijb/vol1/iss1/4>).
7. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;15:615–25.
8. Hernán MA, Brumback B, Robins JM. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Stat Med* 2002;21:1689–709.
9. Robins JM, Hernán MA, Siebert U. Effects of multiple interventions. In: Ezzati M, Lopez AD, Rodgers A, et al, eds. *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors*. Vol II. Geneva, Switzerland: World Health Organization, 2004:2191–230.
10. van der Laan MJ. Causal effect models for intention to treat and realistic individualized treatment rules. (U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 203). Berkeley, CA: Division of Biostatistics, School of Public Health, University of California, Berkeley, 2006. (<http://www.bepress.com/ucbbiostat/paper203>).
11. Robins JM. 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*. Washington, DC: National Center for Health Services Research, US Public Health Service, 1989:113–59.
12. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Commun Stat* 1994;23:2379–412.
13. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health* 2006;60:578–86.
14. Robins JM, Rotnitzky A. Comment on the Bickel and Kwon article, “Inference for semiparametric models: some questions and an answer.” *Stat Sinica* 2001;11:920–36.

15. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 1995;90:106–21.
16. van der Laan M, Robins JM. Unified methods for censored and longitudinal data and causality. New York, NY: Springer Verlag, 2003.
17. van der Laan MJ, Hubbard AE, Jewell NP. Estimation of treatment effects in randomized trials with noncompliance and a dichotomous outcome. (U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 157). Berkeley, CA: Division of Biostatistics, School of Public Health, University of California, Berkeley, 2004. (<http://www.bepress.com/ucbbiostat/paper157>).
18. Orellana L, Rotnitzky A, Robins JM. Generalized marginal structural models for estimating optimal treatment regimes. (Technical report). Boston, MA: Department of Biostatistics, Harvard School of Public Health, 2006.
19. Murphy SA. Optimal dynamic treatment regimes. *J R Stat Soc B* 2003;65:331–66.
20. Robins JM. Optimal structural nested models for optimal sequential decisions. In: Lin DY, Heagerty P, eds. Proceedings of the Second Seattle Symposium on Biostatistics. New York, NY: Springer Publishing Company, 2004.

APPENDIX

Here we prove a number of results mentioned in the main text as well as briefly touch on certain more advanced issues. Our discussion is restricted to structural mean models, that is, models for the conditional mean of a counterfactual outcome.

Estimation of the parameters of marginal structural models (MSMs)

Throughout we use the following notational conventions. Capital letters such as $L(m)$ refer to random variables, that is, a variable which can take on different values for different study subjects. Small letters such as $l(m)$ refer to the possible values of $L(m)$. Overbar variables with a time t in parentheses denote the history of the variable from 0 to t , and overbars without parentheses denote the entire covariate history, that is, $\bar{A}(t) = \{A(0), \dots, A(t)\}$ and $\bar{A} = \bar{A}(K)$. In addition, we use underbars to denote future values of a variable in the following way: $\underline{A}(m, t) = \{A(m), A(m+1), \dots, A(t-1), A(t)\}$ is the A -history from time m through time t and $\underline{A}(m) = \underline{A}(m, K)$ is a subject's treatment history from m to the end of the study. Similarly, we let $\underline{a}(m, t)$ and $\underline{a}(m) = \underline{a}(m, K)$ denote a possible treatment history from m to t and from m to K , respectively. By convention, $a(m) = 0$ denotes either no treatment or a standard treatment at time m . Thus, the history $\underline{a}(m, t) = \underline{0}(m, t)$ stands for the history “withhold treatment from m through t ” or “receive the standard treatment from m through t .”

Let $\mathbf{H}(m) = \{\bar{L}(m), \bar{A}(m-1)\}$ be the entire covariate and treatment history prior to receiving treatment $A(m)$, and let $V(m)$ be a subvector of $\mathbf{H}(m) = \{\bar{L}(m), \bar{A}(m-1)\}$

that is of interest as an effect modifier. We may sometimes choose $V(m)$ to be all of $\mathbf{H}(m)$. Let $Y_{\underline{a}(m)}(t)$ be a subject's counterfactual outcome at time t if the subject had received his observed treatment regime $\bar{A}(m-1)$ up to time m and history $\underline{a}(m)$ from m onwards.

A standard, univariate MSM models the mean of the counterfactual outcome $Y_{\underline{a}(m)}(t)$ at time $t > m$ as a function of the possible treatment histories $\underline{a}(m, t-1)$ from time m to time $t-1$ and the baseline covariates $V(m)$. For example,

$$E[Y_{\underline{a}(m)}(t)|V(m)] = r(t, m, \underline{a}(m, t-1), V(m), \theta^*, \beta^*)$$

and

$$\begin{aligned} r(t, m, \underline{a}(m, t-1), V(m), \theta^*, \beta^*) \\ = \theta_0^* + \theta_1^* V(m) + \beta_0^* \text{dur}[\underline{a}(m, t-1)] \\ + \beta_1^* \text{dur}[\underline{a}(m, t-1)]V(m), \end{aligned}$$

where $\theta^* = (\theta_0^*, \theta_1^*)$ and $\beta^* = (\beta_0^*, \beta_1^*)$ are unknown parameter vectors and

$$\text{dur}[\underline{a}(m, t-1)] = \sum_{j=m}^{t-1} a(j)$$

is the cumulative treatment from the baseline time m through t under regime $\underline{a}(m)$.

Under the assumption of no unmeasured confounders for the effect of the time-varying treatment $A(m)$, the parameters of the univariate MSM can be estimated by weighted least squares with estimated stabilized inverse probability weights depending on the baseline time m and response time t :

$$\widehat{\text{SW}}(m, t) = \frac{\prod_{j=m}^{t-1} \hat{f}[A(j)|\bar{A}(j-1), V(m)]}{\prod_{j=m}^{t-1} \hat{f}[A(j)|\mathbf{H}(j)]}.$$

The parameters of the multivariate model are estimated using a weighted generalized estimating equations program with (m, t) -specific weights $\widehat{\text{SW}}(m, t)$ and with a user-supplied subject-specific working covariance matrix. If one chooses an independence working covariance matrix, the estimate of $E[Y_{\underline{a}(m)}(t)|V(m)]$ based on the univariate MSM will be exactly equal to that based on the multivariate MSM with time-specific parameters. The same holds true for MSMs with multiple baseline times m .

Model incompatibility in history-adjusted MSMs

To explain why history-adjusted MSMs may be incompatible, we will consider a univariate MSM with $t = K + 1$ that allows the effect of the treatment $a(m)$ on $Y(K + 1)$ to differ from the effect of later treatments. It will be helpful to decompose $E[Y_{\underline{a}(m)}(K + 1)|V(m)]$ into the sum of three functions:

1. The conditional mean of $Y(K + 1)$ when treatment is withheld from m onwards:

$$r_0(K + 1, m, V(m)) = E[Y_{(\underline{0}(m), \underline{0}(m+1))}(K + 1)|V(m)].$$

2. The direct effect of treatment $a(m)$ on $Y(K + 1)$ when treatment from $m + 1$ onwards is withheld (i.e., $\underline{a}(m + 1, K) = 0$):

$$r_1(K+1, m, a(m), V(m)) = E[Y_{(a(m), \underline{0}(m+1))}(K+1)|V(m)] \\ - E[Y_{(0(m), \underline{0}(m+1))}(K+1)|V(m)].$$

3. The effects of treatment $\underline{a}(m+1, K)$ from $m+1$ onwards (including effects due to interactions with treatment $a(m)$ at m):

$$r_2(K+1, m, \underline{a}(m, K), V(m)) = E[Y_{\underline{a}(m)}(K+1)|V(m)] \\ - E[Y_{(a(m), \underline{0}(m+1))}(K+1)|V(m)].$$

If we specify parametric models for these three unknown functions, we obtain a model $r(K+1, m, \underline{a}(m, K), V(m), \theta^*, \beta^*)$ for $E[Y_{\underline{a}(m)}(K+1)|V(m)]$. As a concrete example, suppose we specify

1. $r_0(K+1, m, V(m), \theta^*) = \theta_{0,m}^* + \theta_{1,m}^* V(m)$;
2. $r_1(K+1, m, a(m), V(m), \beta^{(1)*}) = \beta_0^{(1)*} a(m) + \beta_1^{(1)*} a(m) \times V(m) + \beta_2^{(1)*} a(m)(K-m)$; and
3. $r_2(K+1, m, \underline{a}(m, K), V(m), \beta^{(2)*}) = \beta_0^{(2)*} \text{dur}[\underline{a}(m+1, K)] + \beta_1^{(2)*} \text{dur}[\underline{a}(m+1, K)|V(m)] + \beta_2^{(2)*} \text{dur}[\underline{a}(m+1, K)](K-m) + \beta_3^{(2)*} a(m) \text{dur}[\underline{a}(m+1, K)]$,

where, by definition, $\text{dur}[\underline{a}(K+1, K)] = 0$. This model assumes that the main effect of treatment at m and of subsequent cumulative treatment $\text{dur}[\underline{a}(m+1, K)]$ are modified by the baseline time m only through the linear term $(K-m)$. The vector $\beta^{(1)*}$ encodes the direct effects of $a(m)$ on $Y(K+1)$ when treatment from $m+1$ onwards is withheld, while the vector $\beta^{(2)*}$ encodes all effects of $\underline{a}(m+1, K)$ on $Y(K+1)$, including its effect due to interactions with $a(m)$ encoded in the parameter $\beta_3^{(2)*}$.

A given treatment regime $g(m) = \{g_m\{\mathbf{h}(m)\}, g_{m+1}\{\mathbf{h}(m+1)\}, \dots, g_K\{\mathbf{h}(K)\}\}$ has a subject follow her observed treatment history up to m and then, at each time $j \geq m$, determines her treatment dose at j by the value of a given function $g_j\{\mathbf{h}(j)\}$ of past treatment and covariate history $\mathbf{h}(j)$. If, for each $j \geq m$, $g_j\{\mathbf{h}(j)\}$ gives the same value $a(j)$ for all past $\mathbf{h}(j)$, we can say that the regime $g(m)$ is nondynamic and write the regime as $\underline{a}(m) = \{a(m), \dots, a(K)\}$, as in the main text. Otherwise, the regime is dynamic. The following is an immediate consequence of theorem 4 in the paper by Robins (12).

Theorem 1. Suppose the sequential randomization assumption holds (i.e., there are no unmeasured confounders) for all m and that all treatments $a(m)$ are coded as nonnegative. Suppose that for each m the effect of $a(m)$ on the mean of $Y(K+1)$ is less than 0 when $\underline{a}(m+1, K) = 0$ —that is, for all m , $\mathbf{H}(m)$:

$$r_1(K+1, m, a(m), \mathbf{H}(m)) < 0.$$

Then

$$r_2(K+1, m, \underline{a}(m, K), \mathbf{H}(m)) < 0$$

and

$$E[Y_{g(m)}|\mathbf{H}(m)] \leq E[Y_{\underline{0}(m)}(K+1)|\mathbf{H}(m)]$$

for any regime $g(m)$.

We now discuss the relevance of theorem 1 for the example given in the text. Suppose all levels $\mathbf{h}(m)$ of $\mathbf{H}(m)$ are coded as nonnegative and $V(m) = \mathbf{H}(m)$. It follows from the theorem that the inverse-probability-weighted estimates $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ in our example are logically inconsistent (in the sense that no actual distribution exists with these parameter values), since all components of $\hat{\beta}^{(1)}$ being negative and all components of $\hat{\beta}^{(2)}$ being positive imply that our estimate $r_1(K+1, m, a(m), V(m), \hat{\beta}^{(1)})$ of $r_1(K+1, m, a(m), \mathbf{H}(m))$ is negative but our estimate $r_2(K+1, m, \underline{a}(m, K), V(m), \hat{\beta}^{(2)})$ of $r_2(K+1, m, \underline{a}(m, K), V(m))$ is positive for all $\mathbf{H}(m)$, which contradicts the above theorem.

Furthermore, the last part of theorem 1 implies, as stated in the text, that our negative estimate of $\hat{\beta}^{(1)}$ means that the regime $\underline{0}(m)$ is the estimated optimal regime. We next verify our claim in the text that components of $\hat{\beta}^{(2)}$ positive and larger in absolute value than those of $\hat{\beta}^{(1)}$ imply that the regime $\underline{1}(m)$, “always take treatment starting at m ,” is estimated to be preferable to the regime $\underline{0}(m)$. Note that

$$E[Y_{\underline{1}(m)}(K+1)|\mathbf{H}(m)] - E[Y_{\underline{0}(m)}(K+1)|\mathbf{H}(m)] \\ = r_1(K+1, m, \underline{1}, \mathbf{H}(m)) + r_2(K+1, m, \underline{1}(m, K), \mathbf{H}(m)).$$

By the above relation between $\hat{\beta}^{(2)}$ and $\hat{\beta}^{(1)}$, our estimate of $r_2(K+1, m, \underline{1}(m, K), \mathbf{H}(m))$ is positive and much larger in absolute value than our estimate of $r_1(K+1, m, \underline{1}, \mathbf{H}(m))$. It then follows that our estimate of

$$E[Y_{\underline{1}(m)}(K+1)|\mathbf{H}(m)] - E[Y_{\underline{0}(m)}(K+1)|\mathbf{H}(m)] \\ = r_1(K+1, m, \underline{1}, \mathbf{H}(m)) + r_2(K+1, m, \underline{1}(m, K), \mathbf{H}(m))$$

is positive.

Finally, we stress that incompatible estimates often pose no difficulty when they cannot result in logically contradictory estimates of substantively important effects. As an example, Robins and Rotnitzky (14) argued that using incompatible models and estimates to construct generalized doubly robust estimators posed no problem, because the models served simply as statistical tools for reducing bias. In contrast, use of incompatible history-adjusted MSMs can be problematic, because they are substantive tools used to estimate treatment effects.

Structural nested models (SNMs) for handling model incompatibility

Standard inverse probability weighting methods require that the positivity assumption $f[a(j)|\mathbf{H}(j)] > 0$ hold for all possible values of $a(j)$ and (essentially) all histories $\mathbf{H}(j)$. Even when the positivity assumption holds, the denominator of $\widehat{SW}(m, K)$,

$$\prod_{j=m}^{K-1} \hat{f}[A_j|\mathbf{H}(j)],$$

can be difficult to model well, can vary greatly between subjects, and can be exceedingly small for certain subjects,

particularly when $K - m$ is large, the treatment $A(j)$ has many levels or is continuous, and/or there exist many continuous covariates in $L(j)$. As a consequence, the few subjects whose weights $\widehat{SW}(m, K)$ are largest may have a huge effect on the analysis, leading to decreased precision, finite-sample bias, and often severe large-sample bias because misspecification of a model for $f[A_j|\mathbf{H}(j)]$ often disproportionately affects the largest weights (15). (Joffe et al. (5) initially proposed a particular type of history-adjusted MSM as a means of partially surmounting the problem of extreme weights for large $K - m$.) Although there exist a number of ways to partially surmount these problems, such as doubly robust estimation, use of various diagnostics, truncation of extreme weights, etc., none is entirely satisfactory.

In contrast, g -estimation of a structural nested mean model $r_1(K + 1, m, a(m), \mathbf{H}(m), \beta^{(1)*})$ for the direct effect $r_1(K + 1, m, a(m), \mathbf{H}(m))$ of treatment $a(m)$ does not require the positivity assumption and is much less affected by $K - m$ being large, the treatment $A(j)$ having many levels or being continuous, and there being many continuous covariates in $L(j)$. First, one does not divide by estimates of $f[A(j)|\mathbf{H}(j)]$, so the problem of extreme weights does not exist. In fact, those subjects who would have the most extreme weights and thus cause the most trouble for inverse probability weighting make a much smaller contribution to the g -estimation analysis, thereby causing little trouble. Second, for continuous or many-leveled $A(j)$'s, one need only model the mean of $A(j)$ given $\mathbf{H}(j)$, a much easier task than modeling the entire density function $f[a(j)|\mathbf{H}(j)]$. Third, even if $K - m$ is large, one can choose not to model the mean of $A(j)$ given $\mathbf{H}(j)$ for large j near K , thereby trading off some loss of precision for better bias control.

Another apparent advantage of estimating $r_1(K + 1, m, a(m), \mathbf{H}(m))$ by g -estimation of an SNM rather than inverse probability weighting estimation of an MSM is that no model for $r_0(K + 1, m, \mathbf{H}(m))$ is required. However, this advantage is only apparent; Robins (1) describes a modification of an MSM, referred to as a “semiparametric regression MSM,” that also does not require a model for $r_0(K + 1, m, \mathbf{H}(m))$ and is fitted by inverse probability weighting.

In our example, we took $V(m)$ to be the entire past $\mathbf{H}(m)$. When $V(m)$ and $\mathbf{H}(m)$ differ, a model for $r_1(K + 1, m, a(m), V(m))$ is referred to as a “marginal structural nested model”; as befits its name, a hybrid of g -estimation and inverse probability weighting estimation is used to estimate the model parameters. (See van der Laan and Robins (16) for details.)

Finally, g -estimation of an SNM, unlike inverse probability weighting estimation of an MSM, has not been possible when the response $Y(t)$ was a dichotomous indicator of disease status, except under the rare disease assumption. Hence, history-adjusted MSMs might be preferred to SNMs for nonrare dichotomous responses. However, recent work by van der Laan et al. (17) and Richardson and Robins (T. Richardson and J. Robins, Harvard School of Public Health, unpublished data) holds the promise that, in the near future, g -estimation of SNMs may be extended to cover nonrare dichotomous responses.

We now describe how, after obtaining a g -estimate $\hat{\beta}^{(1)}$ of the parameter $\beta^{(1)*}$ of an SNM $r_1(K + 1, m, a(m), \mathbf{H}(m); \beta^{(1)*})$, we can use Monte Carlo simulation to estimate $E[Y_{g(m)}(K + 1)]$ for any $g(m)$ without having to model $r_2(K + 1, m, a(m), K, \mathbf{H}(m))$. First we estimate $E[Y_{(0(m))}(K + 1)]$ by the sample average of

$$\begin{aligned} \hat{E}[Y_{(0(m))}(K + 1)] &= Y(K + 1) \\ &\quad - \sum_{j=m}^K r_1(K + 1, j, a(j), \mathbf{H}(j); \hat{\beta}^{(1)}). \end{aligned}$$

Then we estimate $E[Y_{g(m)}(K + 1)]$ as follows.

1. First, for $k = m, \dots, K$, fit a parametric model for $f[l(k)|\bar{a}(k - 1), \bar{l}(k - 1)]$ to the data and let $\hat{f}[l(k)|\bar{a}(k - 1), \bar{l}(k - 1)]$ denote the estimate of $f[l(k)|\bar{a}(k - 1), \bar{l}(k - 1)]$ under the model.
2. Do the following for $v = 1, \dots, V$, with V selected to be very large:
 - a) Choose $\mathbf{h}_v(m) = \bar{l}_v(m), \bar{a}_v(m - 1)$ to be the value of $\mathbf{H}(m)$ for a subject randomly drawn from the n study subjects.
 - b) Recursively for $k = m + 1, \dots, K$, draw $l_v(k)$ from $\hat{f}[l(k)|\bar{a}_v(k - 1), \bar{l}_v(k - 1)]$ with the treatment history from m to $k - 1$ determined by the regime $g(m)$.
 - c) Let $\hat{\Delta}_{g(m), v} = \sum_{j=m}^K r_1(K + 1, j, a_v(j), \mathbf{h}_v(j), \hat{\beta}^{(1)})$.
3. Let $\hat{E}[Y_{g(m)}(K + 1)] = \hat{E}[Y_{(0(m))}(K + 1)] + \sum_{v=1}^V \hat{\Delta}_{g(m), v} / V$ be the estimate of $E[Y_{g(m)}(K + 1)]$.

The above approach is based on theorem 4 in the paper by Robins (12). Alternative approaches to the estimation of $E[Y_{g(m)}(K + 1)]$ for both dynamic and nondynamic regimes, based on other recent extensions of MSMs and SNMs, have been developed by Orellana et al. (18), van der Laan et al. (10), Murphy et al. (19), and Robins (20).

Alternative approaches to handling model incompatibility

We now discuss alternative approaches to handling model incompatibility and consider their possible application to history-adjusted MSMs.

Saturated models. If an incompatible model is saturated, one will never obtain incompatible parameter estimates. Thus, in the context of our linear probability example, if we fit the saturated incompatible model

$$\Pr(D = 1|X) = \alpha_0 + \sum_{k=1}^{100} \alpha_k I_k,$$

where I_k is a dummy variable that takes the value 1 if $X = k$ and 0 otherwise, our estimates of $\Pr(D = 1|X)$ are guaranteed to lie between 0 and 1 (although the associated confidence intervals may contain incompatible values). Unfortunately, because the data in realistic longitudinal studies are sparse and high-dimensional, fitting saturated history-adjusted MSMs is not possible. Therefore, possibly misspecified nonsaturated models must be used.

Replacing models with approximations. All models are incorrect. Van der Laan et al. (6) argue that it is therefore more honest to redefine $\beta^{(1)*}$ and $\beta^{(2)*}$ in the history-adjusted MSM of our example to be the limits of the inverse-probability-weighted estimates $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ as the sample size goes to infinity. They then view $r_1(K + 1, m, a(m), V(m); \beta^{(1)*})$ and $r_2(K + 1, m, \underline{a}(m, K), V(m); \beta^{(2)*})$ as approximations of, rather than models for, $r_1(K + 1, m, a(m), V(m))$ and $r_2(K + 1, m, \underline{a}(m, K), V(m))$. From this point of view, since there are no models, there is no possibility of model or parameter incompatibility. Thus, neither $\beta^{(1)*}$ and $\beta^{(2)*}$ nor $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ can be incompatible.

Our difficulty with this approach is that it does nothing to solve our problem; it simply sweeps the problem under the rug. In the context of our example, a health-care provider remains without a clue as to a reasonable treatment strategy, since she can still deduce from theorem 1 that it is logically impossible for both $r_1(K + 1, m, a(m), V(m); \hat{\beta}^{(1)*})$ to be a good approximation of $r_1(K + 1, m, a(m), V(m); \beta^{(1)*})$ and $r_2(K + 1, m, \underline{a}(m, K), V(m); \hat{\beta}^{(2)*})$ to be a good approximation of $r_2(K + 1, m, \underline{a}(m, K), V(m))$.

Exploiting incompatible models for goodness-of-fit (GOF) testing and model selection. We say that a model indexed by a parameter vector η is correctly specified if there is a true (and therefore compatible) value η^* of η under which the data were generated. All saturated models are correctly specified. In contrast to a saturated model, if one fits a correctly specified incompatible model that is not saturated, one may obtain incompatible parameter estimates; however, a $1 - \alpha$ confidence interval for η^* must include the true compatible parameter vector η^* and, thus, a compatible parameter value with probability at least $1 - \alpha$. Therefore, we can perform a valid (albeit conservative) α -level GOF test of the null hypothesis that an incompatible model is correctly specified by rejecting the null hypothesis whenever a $1 - \alpha$ confidence interval for η^* fails to contain a compatible parameter value η . If the GOF test accepts, we accept the null hypothesis of correct specification, and

the set of compatible parameter values η in the $1 - \alpha$ confidence interval for η^* forms a $1 - \alpha$ confidence set for η^* .

If, as would be the case in our example with $\eta^* = (\beta^{(1)*}, \beta^{(2)*})$, our GOF test rejects, we enlarge our model by increasing the dimension of η^* —for example, by adding quadratic interactions with time, $\beta_3^{(1)*} a(m)(K - m)^2$ and $\beta_4^{(2)*} \text{dur}[\underline{a}(m + 1, K)](K - m)^2$ —and then testing whether the enlarged model fits. If not, we continue enlarging until we finally have a model that fits, and we report the set of compatible values of the enlarged parameter η contained in the $1 - \alpha$ confidence interval for the enlarged η^* as a $1 - \alpha$ confidence set for η^* . The actual coverage of these intervals would not be $1 - \alpha$, but appropriate corrections could be worked out. Furthermore, one needs an algorithm for finding the set η of compatible values in a given $1 - \alpha$ confidence interval for η^* , which is a highly nontrivial problem. In addition, the power properties of this procedure are almost certainly poor.

With much additional work, it is conceivable that this GOF-testing-based model selection strategy might someday become, in certain settings, a viable alternative to the strategy of using SNMs rather than history-adjusted MSMs. A naive reader might think the strategy based on GOF testing even has certain advantages over the use of SNMs, since, if the model $r_1(K + 1, m, a(m), V(m); \beta^{(1)*})$ is badly misspecified, the GOF approach might detect such misspecification while the most straightforward use of g -estimation will not.

However, if one really wishes to perform a GOF test of the model $r_1(K + 1, m, a(m), V(m); \beta^{(1)*})$ with $V(m) = \mathbf{H}(m)$, GOF tests based on g -estimation of enlargements of the model should be more efficacious and powerful than the above inverse-probability-weighting-based GOF test of compatibility of the model $r_1(K + 1, m, a(m), V(m); \beta^{(1)*})$ with the model $r_2(K + 1, m, \underline{a}(m, K), V(m); \beta^{(2)*})$, since the latter model may itself be badly misspecified. Thus, we are skeptical that the use of incompatible models for GOF testing and model selection will prove beneficial.