The International Journal of Biostatistics

Volume 6, Issue 2

2010

Article 13

CAUSAL INFERENCE

Accuracy of Conventional and Marginal Structural Cox Model Estimators: A Simulation Study

Yongling Xiao, *McGill University*

Michal Abrahamowicz, *McGill University & Montreal
General Hospital*

Erica E. M. Moodie, *McGill University*

Recommended Citation:

Xiao, Yongling; Abrahamowicz, Michal; and Moodie, Erica E. M. (2010) "Accuracy of Conventional and Marginal Structural Cox Model Estimators: A Simulation Study," *The International Journal of Biostatistics*: Vol. 6: Iss. 2, Article 13.

DOI: 10.2202/1557-4679.1208

Accuracy of Conventional and Marginal Structural Cox Model Estimators: A Simulation Study

Yongling Xiao, Michal Abrahamowicz, and Erica E. M. Moodie

Abstract

Marginal structural models (MSM) provide a powerful tool to control for confounding by a time-dependent covariate without inappropriately adjusting for its role as a variable affected by treatment (Hernán et al., 2000). In this paper, we demonstrate that it is possible to fit a marginal structural Cox model directly, rather than the typical approach of using pooled logistic regression, using the weighted Cox proportional hazards function that has been implemented in standard software. To evaluate the performance of the marginal structural Cox model directly via inverse probability of treatment weighting, we conducted several simulation studies based on two data-generating models: one which replicates the simulations of Young et al. (2009) and an additional, more clinically plausible approach which mimics survival data with time-dependent confounders and time-varying treatment. Using the simulations, we illustrate the limitations of the conventional time-dependent Cox model and the MSM fitted via pooled logistic regression. Furthermore, we propose two novel normalized weights with the goal of reducing the MSM estimators' variability. The performance of the normalized weights is evaluated alongside the usual unstabilized and stabilized weights.

KEYWORDS: survival analysis, longitudinal data, marginal structural model, inverse probability weighting, time-dependent confounder, mediator

Author Notes: We wish to thank the editor and an anonymous referee for their helpful suggestions. Yongling Xiao holds a doctoral research scholarship from the Fonds de Recherche sur la Nature et les Technologies du Québec (FQRNT). Michal Abrahamowicz is a James McGill Professor at McGill University. Erica Moodie acknowledges the support of a Natural Sciences and Engineering Research Council of Canada (NSERC) University Faculty Award. This work was partially supported by NSERC grants (#228203 and #331271) and CIHR grant (#MOP-8127).

1 Introduction

In survival analysis, time-dependent predictors of survival that are also predictors of subsequent treatment are known as time-dependent confounders. Often there exists a time-dependent confounder that is also a mediating variable, i.e. treatment modifies the subsequent values of such a variable, which in turn alters the risk of a negative outcome. If one then wishes to estimate the direct and indirect, or mediated, effects of a time-varying sequence of treatments on survival, the causal structure of the data implies that the analysis should not adjust for this variable. Indeed, adjusting for a mediating variable, which is on the causal pathway, could bias the estimation of the indirect or total treatment effects (Robins et al., 2000). However, if the analysis does not adjust for this variable, the estimator of the current treatment effect will also be biased, due to confounding. Thus, the conventional multivariable Cox model yields biased estimator of the total causal effect of treatment, even in the absence of unmeasured confounders and model misspecification, in the situation where time-varying confounders act also as mediating variables.

Marginal structural models (MSMs) have been proposed to estimate the causal effect of a time-varying treatment, controlling for confounding by a time-dependent covariate without inappropriately adjusting for its role as a variable affected by treatment (Robins, 1997, 1999a,b; Robins et al., 2000; Robins and Hernán, 2008). Hernán et al. (2000; 2001) introduced the marginal structural Cox proportional hazards model for survival analysis. Instead of adjusting for the time-dependent confounder, in a marginal structural Cox model, inverse-probability-of-treatment weighting (IPTW) is used to yield a consistent estimate of the causal effect of a time-varying treatment on the hazard. This is accomplished by first calculating weights, one for each subject at each time interval, which are functionally related to the decision to take a particular treatment (the decision may include starting, continuing, or interrupting a therapeutic option). These time-dependent weights are calculated as the inverse of the estimated probability of receiving the observed treatment, conditional on past confounder history, treatment history and baseline measurements. A weighted time-dependent regression model is then fitted with the contribution of each subject to a risk set at a given time weighted by the corresponding weight (Hernán et al., 2000).

Hernán et al. (2000) proposed the use of a weighted pooled logistic regression based on a discretization of the survival time to approximate the weighted continuous-time Cox model. To date, virtually all applications of marginal structural models in time-to-event analyses have employed the pooled logistic approximation to estimate the MSM parameters (see, for example, Choi et al.

2002; Cook et al. 2002; Cole et al. 2003; Bryan et al. 2004; Sterne et al. 2005; Westreich et al. 2009), even though this approximation has been found to lead to biased estimators of the treatment effect when events are frequent (Young et al., 2009). To the best of our knowledge, no systematic comparisons of the performance of the weighted pooled logistic and the weighted Cox models in the MSM analyses of time-to-event data with time-varying treatments and time-dependent confounder/mediating variables have been reported in the literature.

Another issue that requires further consideration and evaluation is the calculation of the weights in a time-varying setting. Two different weights were considered by Hernán et al. (2000): unstabilized weights and stabilized weights. For a given subject at a specific time t , the two weights have the same denominator, which is the product of the consecutive probabilities, calculated for all times before and up to t when a treatment was assigned, that a subject received his own observed treatment at a particular time, given baseline covariates and the history of treatments and time-dependent confounders. However, the numerators of the two weights differ. The numerator of a stabilized weight is the product of the probability of receiving the observed treatment conditional on baseline covariates and past treatments (but not time-dependent confounders), while the numerator for an unstabilized weight is 1. The stabilized weight is most commonly used and issues relating to its use such as truncation to reduce variability have been considered by Cole and Hernán (2008).

The goal of this paper is to investigate several issues related to the fitting of marginal structural Cox models. We show how a marginal structural Cox model, using exact event times measured on a continuous time scale, can be implemented with standard software, as an alternative to the weighted pooled logistic approximation, which uses discrete event times. We then conduct a series of simulations, in which event times are generated conditional on both (i) time-varying treatment and (ii) mediating time-dependent confounder, assuming different data-generating mechanisms. We then use the simulated data to evaluate and compare the performance of the marginal structural Cox model estimated via IPTW time-dependent Cox regression versus the marginal structural Cox model estimated via IPTW pooled logistic regression in estimating the causal effect of treatment. We also compare the performance of the unstabilized and stabilized weights. In addition, two novel normalized weights are proposed, and are compared with the two conventional weights.

The paper is organized as follows. In Section 2, we describe the design and methods of simulations, define the two widely used IPTW weights, and propose two new weights. Section 3 describes the design and the methods of the

simulation studies. Section 4 summarizes the results of different simulations and compares the fitting of the marginal structural Cox model directly with the pooled logistic regression approximation, as well as the different weighting methods for the direct fitting of a marginal structural Cox model. We conclude the paper with the discussion of our results, the limitations of our study, and suggestions for future research, in Section 5.

2 Marginal Structural Cox Modelling

2.1 Notation for and definition of the marginal structural Cox model

We consider a longitudinal study, in which N subjects were followed at regular time intervals since their entry into the cohort (time 0). Let T_i denote the observed survival time of subject i , and let $A_i(t)$ be a binary indicator of the treatment assigned to subject i at time t ($A_i(t) = 1$ if a subject received treatment and $A_i(t) = 0$ otherwise). We denote the vector of baseline covariates by V_i and the time dependent covariate by $L_i(t)$. We use over-bar to represent a covariate history, thus $\bar{A}(t)$ indicates the treatment history up to t , and $\bar{L}(t)$ indicates the history of time-dependent confounders.

Let $T_{\bar{a}}$ denote the random variable representing a subject's survival time had he followed the treatment history $\bar{a} = \{a(t); t \geq 0\}$. For each $\bar{a}$, the marginal structural Cox proportional hazards model can be specified as in Hernán et al. (2001):

$$\lambda_{T_{\bar{a}}}(t|V) = \lambda_0(t) \exp(\beta_1 f(\bar{a}(t)) + \beta_2 V) \quad (1)$$

where $\lambda_{T_{\bar{a}}}(t|V)$ is the hazard at time t among subjects with the baseline covariates V had, contrary to fact, all subjects followed the treatment history $\bar{a}$, β s are unknown parameters, $\lambda_0(t)$ is the unspecified baseline hazard at time t , corresponding to $V = 0$ and $\bar{a}(t) = \bar{0}$, and $f(\cdot)$ is an analyst-defined function of the treatment history. For example, Hernán et al. (2000) specified $\lambda_{T_{\bar{a}}}(t|V) = \lambda_0(t) \exp(\beta_1 a(t) + \beta_2 V)$, i.e. the MSM is a function of current treatment only.

Robins (1997) has shown that the causal parameter β_1 can be consistently estimated by using inverse-probability-of-treatment weighting to obtain an estimator $\hat{\beta}_1$, under the assumption of no unmeasured confounders given the measured time-dependent risk factor $L(t)$.

2.2 Computation of the IPTW marginal structural Cox model estimator

We propose to implement the estimation of the marginal structural Cox model using standard statistical software for fitting Cox model with time-dependent covariates. This can be achieved as follows: one must first create an augmented dataset, in which each person-visit ($t = 1, \dots, m_i$) corresponds to one row, which contains the time-invariant values of the baseline covariates, the updated values of the current binary treatment indicator $A(t)$, and the current time-dependent confounder $L(t)$. Next, the observed values of the time-dependent treatment indicator are used to estimate the (possibly multiple) logistic models for predicting treatment choices to create time-dependent weights for person-visits, as detailed in Section 2.3. Finally, standard software is used to fit the marginal structural Cox model directly to the augmented data, by specifying the column that contains the values of the time-dependent weights.

2.3 Estimation of the weights for a MSM

To fit a MSM, either unstabilized (Hernán et al., 2000) or stabilized weights (Robins et al., 2000) are commonly used for estimation via IPTW. In the context of time-varying treatments, which are the focus of our study, the IPTW weights are time-dependent, i.e. the same subject is assigned different weights at different visits. Specifically, the IPTW weight for visit j , $1 < j < m$, is based on the overall probability of the subject receiving his or her own observed sequence of treatments for all previous visits $A_i(1)$ to $A_i(j)$, i.e. the product of the visit-specific probabilities of the observed treatments. In particular, the unstabilized weight, $w_i(t)$, is calculated as the inverse of the estimated probability that a subject i received his own observed treatment, given his baseline covariates, past treatments, and confounder history:

$$w_i(t) = \prod_{j=1}^{m(t)} \frac{1}{P[A(j) = a_i(j) | \bar{A}(j-1) = \bar{a}_i(j-1), \bar{L}(j) = \bar{l}_i(j), V = v_i]} \quad (2)$$

where $m(t)$ is the total number of visits up to t , $\bar{A}(j)$ represents a vector of treatment history up to visit j , $\bar{L}(j)$ represents the history of the confounder up to visit j , and V is the vector of baseline covariates.

The stabilized weights, $w_i^{(s)}(t)$, are a modification of the unstabilized weights, in which the 1 in the numerator of $w_i(t)$ is replaced with the estimated probability of a subject receiving his own observed treatment conditional on his

treatment history and baseline covariates but not on $L(j)$:

$$w_i^{(s)}(t) = \prod_{j=1}^{m(t)} \frac{P[A(j) = a_i(j) | \bar{A}(j-1) = \bar{a}_i(j-1), V = v_i]}{P[A(j) = a_i(j) | \bar{A}(j-1) = \bar{a}_i(j-1), \bar{L}(j) = \bar{l}_i(j), V = v_i]} \quad (3)$$

By accounting for both treatment history and baseline covariates in *both* the numerator and the denominator, the stabilized weight reflects an *incremental* effect of the time-varying confounder on the current treatment choice, over and above the other determinants of the treatment.

The unstabilized time-dependent weights in (2) are forced to increase, often significantly, from one visit to another, as with each additional visit the denominator is multiplied by another positive number less than 1. This implies that the later observations receive systematically higher weights than earlier ones and leads to highly variable weights. In contrast, in the case of the stabilized weights, both the numerator and denominator are the product of $m(t)$ consecutive probabilities and, thus, tend to decrease with each increasing visit. Because the denominator of equation (3) accounts for the time-dependent covariate $L(j)$, usually a powerful determinant of treatment choice, the consecutive probabilities in the denominator should be, on average, higher than the corresponding probabilities in the numerator and, thus, the resulting stabilized weights will tend to decrease over time. The reduced variability of the stabilized weights leads to often remarkably more efficient estimators of treatment effects. We postulated that greater control of the variability of the estimated weights would lead to even more efficient estimators.

To reduce variability of the weights in each risk set over time, we consider normalizing both types of weights. Specifically, we define the normalized unstabilized weight, $w_i^{(n)}(t)$, as:

$$w_i^{(n)}(t) = \frac{w_i(t)N(t)}{\sum_{i \in R(t)} w_i(t)} \quad (4)$$

where $R(t)$ is the risk set at t and $N(t)$ is the total number of subjects in the risk set. The normalized stabilized weights $w_i^{(ns)}(t)$ are calculated by applying the same normalization to the stabilized weights in (3). Normalization in equation (4) ensures that the mean of the individual weights for each visit equals 1, as in conventional unweighted maximum (partial) likelihood estimation. This approach has the common-sense appeal of conserving the classic Cox model property of equal weighting of observations across time and also may be expected to reduce sampling variability of weights and, hence, of the resulting treatment effect estimator.

Table 1: Normalized inverse-probability-of-treatment weights $w_i^{(n)}$ and composition of the pseudo-population in a point-treatment study (Robins et al., 2000).

L	A	Y	N	$P(A L)$	w	$w^{(n)1}$	$pseudoN$
1	1	1	108	0.9	1.11	0.56	60
1	1	0	252	0.9	1.11	0.56	140
1	0	1	24	0.1	10	5	120
1	0	0	16	0.1	10	5	80
0	1	1	20	0.5	2	1	20
0	1	0	30	0.5	2	1	30
0	0	1	40	0.5	2	1	40
0	0	0	10	0.5	2	1	10

¹ The normalized weight for stratum i , $w_i^{(n)}$, is calculated as $w_i^{(n)} = w_i \sum_{i=1, \dots, m} N / \sum_{i=1, \dots, m} w_i N$, where i is classified by $\{L, A, Y\}$.

Robins et al. (2000) illustrate the unbiasedness of the IPTW method of estimation. We borrow the same example here to illustrate how the proposed normalized weights in (4) also yield an unbiased estimate of a causal parameter. Specifically, in Table 1, where distributions of L , A , Y and N are the same as in Table A3 of (Robins et al., 2000), we show (i) the calculation of the normalized weights and (ii) the corresponding generated pseudo-population. We then stratify the resulting pseudo-population (from the last column of Table 1) by the confounder L in Table 2, which shows that L and A are no longer associated with each other in this pseudo-population. Indeed, by regrouping the data from Table 2 across both levels of L , we obtain the crude risk difference of $80/250 - 160/250 = -0.32$, which is equal to the causal risk difference reported by Robins et al. (2000).

3 Simulation Studies: Design and Analysis

To evaluate the performance of the marginal structural Cox model estimated via IPTW time-dependent Cox regression for estimating the parameters of a marginal structural Cox model, we performed several simulation studies based on two data-generating models. We began by replicating the simulations of Young et al. (2009), in which the authors conducted simulations to

Table 2: Pseudo-population created by normalized inverse-probability-of-treatment weights stratified by the confounder L .

	$L = 1$		$L = 0$	
	A=1	A=0	A=1	A=0
Y=1	60	120	20	40
Y=0	140	80	30	10
Total	200	200	50	50

assess three classes of structural models (including the marginal structural Cox model) for the effect of a time-varying exposure on survival. We utilized the same data-generating models to compare the estimators of the marginal structural Cox model parameter, estimated by two different methods: IPTW time-dependent Cox regression versus IPTW pooled logistic regression. Additionally, we performed simulations in the context of a cohort of HIV-positive individuals experiencing treatment (and treatment interruptions) over a five year period.

3.1 Replication of the simulations by Young and colleagues

At the initial stage of simulations, we utilized the data generation methods from a recent study by Young et al. (2009). Specifically, following Young and colleagues, we simulated a longitudinal data with n subjects and p observation times. Let T be the failure time, Y_m be the indicator for the failure by time m (1=yes, 0=no), A_m be a binary treatment during the interval $(m, m + 1)$, and L_m be the time-varying confounder measured at the start of interval $[m, m + 1)$. T_0 was defined as the counterfactual survival time under the never-treated regime. The survival data with time-varying treatment was then generated using the following procedure, based on the data generation mechanism proposed by Young et al. (2009):

Step 1: Generate T_0 from an Exponential distribution with a small $\lambda_0 = 0.01$, implying a rare outcome (these are the two key assumptions for the algorithm to be valid).

Define $L_{-1} = A_{-1} = Y_0 = 0$. Then for each $m \in [0, 9]$, do Steps 2 to 4:

Step 2: Generate the confounder L_m from a binomial distribution with

$$\text{logit}[P(L_m = 1 | \bar{L}_{m-1}, \bar{A}_{m-1}, Y_m = 0)] = \beta_0 + \beta_1 I[T_0 < c] + \beta_2 A_{m-1} + \beta_3 L_{m-1}$$

where c is a constant.

Step 3: Generate the treatment A_m from a binomial distribution with

$$\text{logit}[P(A_m = 1 | \bar{L}_{m-1}, \bar{A}_{m-1}, Y_m = 0)] = \alpha_0 + \alpha_1 L_m + \alpha_2 L_{m-1} + \alpha_3 A_{m-1}.$$

Step 4: Generate Y_{m+1} , and possibly T , based on the following:

- If $T_0 > \int_0^{m+1} \exp(\psi A_j) dj$, then $Y_{m+1} = 0$,
- Otherwise, if $T_0 \leq \int_0^{m+1} \exp(\psi A_j) dj$, then $Y_{m+1} = 1$ and $T \in (m, m+1]$ with $T = m + (T_0 - \int_0^m \exp(\psi A_j) dj) \exp(-\psi A_m)$.

As in Young et al. (2009), we simulated 1,000 independent random samples, each with 2,500 subjects and 10 observation times. When analyzing the simulated data, we compared the bias, variance and root mean squared error (RMSE) of the treatment effect estimates obtained with two methods of estimating the marginal structural Cox model. The first method was the weighted pooled logistic regression employed in the original simulation study reported by Young et al. (2009). The second method was the weighted time-dependent Cox model described in Section 2.2 with the stabilized weights defined in equation (3).

3.2 Simulations for a cohort study of HIV-positive patients

In the simulations by Young et al. (2009), the association of the confounder with the survival time was dichotomized using an identity function $I(T_0 < c)$, and kept constant over time. The assumptions were convenient in simplifying the data generation but their clinical plausibility is uncertain. Therefore, in our main simulations, we considered a more complex, and arguably more clinically plausible, causal data structure.

Specifically, we simulated a prospective study of a hypothetical cohort of N HIV-positive patients. Using the notation of Section 2.1, for the i^{th} ($i = 1, \dots, N$) patient, the survival time from the start of the follow-up to an AIDS-defining event was represented by T_i , measured in years. Patients were assumed to be evaluated at regularly spaced visits every 6 months, at which the decisions to initiate, continue, or interrupt the treatment was made. At each

visit j ($j = 1, \dots, m$), the time-varying binary treatment, HAART (Highly Active Anti-Retroviral Therapy), was denoted by $A_i(j)$ ($A_i(j)=1$ if treatment was received and $A_i(j)=0$ otherwise), and a time-dependent confounder, CD4 cell count, was denoted by $L_i(j)$. We assumed that the only baseline covariate was the pre-treatment value of the confounder $L_i(1)$. Throughout, the time-dependent covariate $L_i(j)$ was assumed to act as both a confounder and a mediator for the treatment effect on survival. This assumption is consistent with the role of disease progression markers, such as updated CD4 cell count in HIV-positive individuals, which (i) is associated with mortality hazard, (ii) affects current decisions regarding treatment initiation, interruption or continuation, and (iii) partly reflects the effects of past treatments (Graham et al., 1992; Kinloch-de Loes et al., 1995).

3.2.1 Data generation

To simulate a clinically plausible study while maintaining simplicity of exposition, we assumed that each patient i , $i = 1, \dots, N$, in this cohort was followed for up to 5 years and had a clinical follow-up visit every 6 months. At each visit j , $j = 1, \dots, m$, where $j = 1$ corresponds to the baseline visit at the entry into the cohort ($T = 0$), the current value of the quantitative time-dependent confounder $L_i(j)$ was measured for each subject i . Current treatment assignment $A_i(j)$ was then determined, depending on the current $L_i(j)$ value and the previous treatment $A_i(j - 1)$. We assumed that both $A_i(j)$ and $L_i(j)$ remained constant during the subsequent 6 months between visits j and $(j + 1)$. This implied that all the patients were perfectly compliant with their most recent treatment assignments across the entire 5 year follow-up, and that the time-dependent confounder was a step function of time; both implications are rather unrealistic but expected not to affect the comparisons between alternative models.

Generating the time-varying covariate and treatment

For subject i , $i = 1, \dots, N$, we first generated the baseline covariate $L_i(1) \sim \text{Lognormal}(6, 1)$. The conditional probability of being treated at visit j , $j = 1, \dots, m$, was given by

$$\begin{aligned} \text{logit}(P[A_i(j)|A_i(j-1), L_i(j)]) \\ = 3.623 - 2.605I[L_i(j) > 500] - 0.022(L_i(j) - 200) \\ + 0.009(L_i(j) - 200)I[L_i(j) > 500] + 0.405A_i(j-1) \end{aligned} \quad (5)$$

where we set $A_i(0)=0$ for all i , and defined $I[L_i(j) > 500]$ to take the value 1 if $L_i(j) > 500$ and the value 0 otherwise. Equation (5) reflects the clinically

plausible assumptions that for patients with low current CD4 count ($L_i(j) < 500$), the probability of being treated would be higher and would depend less critically on the exact value of $L_i(j)$.

The subsequent values of $L_i(j)$, $j = 2, \dots, m$, were then assumed to follow a linear function of previous value $L_i(j-1)$, most recent treatment $A_i(j-1)$, and subject-specific rate of natural decay in CD4 count over time Δ_i . The parameter Δ_i was deemed to represent the yearly decline in the i^{th} subject's CD4 count expected in those between-visits time intervals when the subject was *not* treated, and was generated from a Uniform distribution, with $\Delta_i \sim U[-80, -5]$. Specifically, $L_i(j)$ for $2 < j < m$ was generated from the following multivariable linear regression model:

$$L_i(j) = L_i(j-1) + 70A_i(j-1) + \Delta_i + \epsilon_i(j) \quad (6)$$

where $\epsilon_i(j) \sim N(0; \sigma = 3)$.

Generating the outcomes

We generated the individual subjects' event times, conditional on time-varying values of both treatment $A_i(j)$ and covariate $L_i(j)$. Because a subject's hazard varied from one 6-month interval to another, reflecting variation over time in both $A_i(j)$ and $L_i(j)$, we first generated separate potential event times for each interval. The interval survival time t^* , measured from the beginning of a given interval between two adjacent visits, was generated using the standard inversion method from the exponential distribution with the individual interval-specific hazard rate calculated assuming the following proportional hazard model with

$$\lambda_{i,j}[t^*|A_i(j), L_i(j)] = 0.12 \exp[\theta_1 L_i(j) + \theta_2 A_i(j)] \quad (7)$$

where, in our main simulated scenario, the log hazard ratios were set to $\theta_1 = -0.6931$ and $\theta_2 = -0.0016$. All generated values $t^* > 0.5$ years were ignored, i.e. interpreted as the evidence that the subject remained event-free until the end of the respective 6-month interval. Accordingly, if all of the m consecutive t^* values generated for a given subject exceeded 6 months, the subject was considered alive until the end of the 5-year follow-up. Otherwise, we identified the earliest period $1 < k < m$ when the corresponding time did not exceed 0.5 years. The subject was then assumed to have had the event at the follow-up time calculated as $T_i = 0.5(k-1) + t_i^*(k)$ in years since the baseline visit.

The censoring time C_i , from time 0 to the time of potential non-informative right censoring, which was assumed to represent the planned end of follow-up, was generated from a Uniform $U[0, 40]$ years. The observed follow-up time for

the i^{th} subject was then defined as $t_i = \min(T_i, C_i, 5)$ years. The status of subject i at time t was determined as a failure/event ($\delta_i = 1$) if $t_i = T_i$, or censored ($\delta_i = 0$) if $t_i = \min(C_i, 5)$ years. The average censoring rate for the main scenario was approximately 60%.

We generated 2,000 independent random samples, with $N=2,000$, using a maximum follow-up time of $m=10$ visits, or 5 years.

Equation (7) above implies that we generate the event times from a conditional model, in the sense that it specifies the treatment effect as being conditional on the time-dependent covariate $L(j)$. This approach does lead to one complication when using simulated data to evaluate the accuracy of the marginal estimators of the causal effects of the treatment: due to the non-collapsibility of the hazard ratio (see for example Gail 1986; Greenland et al. 1999; Ritz and Spiegelman 2004), the true parameter values of the marginal models may not be easy to obtain analytically. However, in situations such as the simulations we performed, where the event rate in any interval is small (on average, about 4% in each of the ten between-visits intervals), the difference between the marginal and conditional parameters is expected to be negligible, so that non-collapsibility of the hazard ratios should not have any marked impact on our results. To verify this conjecture, we carried out a numerical experiment by generating a very large cohort ($N=50,000$), using the same methods and assumptions as described above with the only exception that the treatment was assigned independently of covariates at every interval, so that $L(j)$ was not a confounder, though it remained a predictor of survival. We then fitted two alternative unweighted Cox models to these data: the conditional model that adjusted for $L(j)$ and the marginal model that did not include $L(j)$. As expected, both models yielded very similar values of the estimated effects of current treatment $A(j)$ (log hazard ratios of 0.668 vs 0.666). This further confirmed that non-collapsibility should not affect materially the results of our simulations.

3.2.2 Analysis of the simulated data

The main objective of the analysis of the simulated data was to compare the accuracy of the treatment effect estimates obtained with alternative versions of the conventional unweighted Cox model with the marginal structural Cox models estimated with different weighting schemes. This involved assuming collapsibility and assessing separately the two effects of treatment, implied by the assumptions underlying our data generation procedures (see Section 3.2.1). As equations (6) and (7) indicate, there are two components to the total treatment effect: (i) the direct effect of current treatment $A_i(j)$, and

(ii) the indirect effect of previous treatment $A_i(j-1)$, mediated through its impact on $L(j)$. The true log hazard ratio for the indirect treatment effect can be calculated by substituting $L(j)$ in equation (7) with equation (6), so that the true hazard $\lambda_{i,j}(t^*)$ becomes now a function of the treatment at the previous interval $A_i(j-1)$, and the corresponding coefficient is equal to $70\theta_1$.

Each simulated sample was analyzed assuming seven different models. All models included the two time-dependent binary indicator variables related to treatment: the binary indicator of current treatment $A(j)$ and that of previous treatment $A(j-1)$. The models differed in how they accounted for the time-dependent covariate $L(j)$. Models 1 and 2 were the conventional unweighted Cox models. Model 1 included the time-dependent covariate $L(j)$ and, thus, adjusted for its most current value, while Model 2 adjusted for the baseline covariate value $L(1)$ only.

Models 3 to 7 all relied on the following marginal structural Cox model, with adjustment for the baseline covariate only:

$$\lambda_{i,j}[t|\bar{A}_i(j), L_i(1)] = \lambda_0(t) \exp[\beta_1 L_i(1) + \beta_2 A_i(j) + \beta_3 A_i(j-1)] \quad (8)$$

MSM models 3 to 6 were estimated using the weighted time-dependent Cox model, described in Section 2.3, with four different types of time-dependent weights: i) unstabilized $w_i(t)$, (ii) normalized (unstabilized) $w_i^{(n)}(t)$, (iii) stabilized $w_i^{(s)}(t)$, and (iv) stabilized normalized $w_i^{(ns)}(t)$ weights, respectively. MSM model 7 was estimated using weighted pooled logistic regression with the stabilized (non-normalized) weights $w_i^{(s)}(t)$.

Because weights varied over visits, we first generated an augmented dataset, in which each person-visit corresponded to one row. Then, we used the equations defined in Section 2.3 to calculate the respective weights for each row. Specifically, we used a logistic model on the augmented data, pooled across all person-visits, to estimate the probability of treatment at visit j , which was used in the denominator of all weights. The binary outcome was the observed treatment at visit j , $A_i(j)$, and the two independent variables were included: the current value of the confounder $L_i(j)$ and the indicator of the previous treatment $A_i(j-1)$. To calculate the numerator of the stabilized weights, we estimated a simpler logistic model, in which the probability of receiving the observed treatment $A_i(j)$ was modeled as a function of $A_i(j-1)$ only.

Once the weights were calculated, we fitted the marginal structural Cox model estimated via IPTW time-dependent Cox regression (Models 3 to 6) or weighted pooled logistic regression (Model 7) directly to the augmented data. Because the use of weights induces within-subject correlation, we used robust variance estimators to provide more conservative estimation of the standard

errors (SEs). We implemented this analysis by using the `weight` and `cluster` arguments in the (time-dependent) `coxph` procedure in R, in the `survival` library (R v2.8.0, 2008).

With the generated 2,000 samples, the bias of the estimators of $A(j)$ and $A(j-1)$ was estimated as the mean difference between the 2,000 estimates and the corresponding true value of the causal parameters. To compare the numerical stability of the estimates obtained with different models, we estimated the empirical standard deviations (SDs) of the 2,000 estimates. Root mean squared errors were then calculated as the square root of the average squared difference between the estimates and the true causal parameters, i.e. of the sum of squared bias and variance.

3.3 Sensitivity analyses

We performed several sensitivity analyses to assess how the results change with different degrees of confounding by varying the current treatment effect or the effect of the confounder on the hazard. We also investigated the impact of sample size ($N=500$, 2,000, and 4,000) on the performance of different models. In an additional sensitivity analysis, we estimated a separate logistic model to predict treatment choice at each of the m visits, instead of estimating a single model using the data pooled from all visits. As the final results were very similar for both methods of modelling treatment choices (data not shown), we present only results based on the pooled model.

4 Results

4.1 Simulations using Young’s data-generation approach

We replicated the simulation study reported by Young et al. (2009) (see Section 3.1 for details). When analyzing the simulated data, in addition to using the approximation based on the pooled logistic regression employed by Young et al. (2009), we estimated the marginal structural Cox model parameter using the weighted time-dependent Cox regression approach outlined in Section 2.2.

The results obtained using the two estimation methods are compared in Table 3. Our results confirmed that for a non-null causal effect (i.e. $\psi \neq 0$), the pooled logistic regression model yielded an unbiased estimate (3rd column) of the marginal structural Cox model parameters when the rare disease condition held ($\lambda_0 = 0.01$), but gave slightly yet systematically biased estimates (7th column) when the rare disease condition was violated ($\lambda_0 = 0.1$). Note that

the bias was statistically significant when the 95% confidence intervals for the bias excluded 0. This is consistent with the original findings reported by Young et al. (2009). In contrast, the IPTW time-dependent Cox regression model always gave unbiased estimates under both rare and common disease conditions. The two estimation methods gave similar SDs, and since all the biases were relatively small, the RMSEs were quite close to the SDs (Table 3).

4.2 The HIV patient cohort simulations

4.2.1 Results for the main simulation

First, we compared the distributions of the weights obtained with different MSMs for a randomly selected sample. As expected, the unstabilized weights showed a dramatic increase over time (mean weight increased from 1.7 at visit 1 to 192.4 at visit 10) and very high variance (SD increased from 6.1 to 1115.1). In contrast, although the distribution of the stabilized weights depends on the numerator used, in general, the mean of the stabilized weights decreased gradually with increasing follow-up time (from 0.9 at the first visit to 0.2 at the last visit in the selected sample), while their standard deviations did not change systematically but were much lower than for the unstabilized weights (between 0.8 and 4.7 depending on the visit). Normalization rescaled the weights proportionally within each visit, thus ensuring the mean weight was equal to 1 at each visit. As with the stabilized weights, the SDs of normalized weights did not exhibit systematic change over follow-up time, but with much lower values and narrower ranges across visits. For the selected sample, the ranges of SDs were 3.6-16.0 for the normalized weights and 2.4-8.7 for the normalized stabilized weights.

Table 4 summarizes the results from 2,000 simulated samples: the left half of the table refers to the direct effect of the current treatment $A(j)$, while the right half to the indirect effect of the treatment at the previous visit $A(j-1)$. The mean number of events per sample was about 780 (60% censoring). The results reflect the assumptions underlying our data-generating procedures, according to which the entire effect of past treatments on the outcome was mediated through the changes in the time-varying confounder. Conventional unweighted Model 1 adjusted for updated values of the time-varying confounder $L(j)$ of the current treatment and, thus, produced an unbiased estimate of the effect of current treatment $A(j)$ but practically eliminated the effect of previous treatment $A(j-1)$, as reflected in about 100% relative bias toward 0 (1st row of Table 4). In contrast, conventional Model 2 (2nd row), which adjusted for the baseline value of the confounder $L(1)$ but not for the current value

Table 3: Comparison of the performance of the marginal structural Cox model estimated via IPTW time-dependent Cox regression versus pooled logistic regression using Young’s data generating mechanism for the marginal structural Cox model based on 1,000 samples, with 2,500 subjects and 10 observation times in each sample.

True	Cox MSM	$\lambda_0 = 0.01$ (rare disease)				$\lambda_0 = 0.1$ (common disease)			
		Bias (95% CI)	Relative Bias(%)	SD	RMSE	Bias (95% CI)	Relative Bias(%)	SD	RMSE
0.3	Weighted Cox	0.006 (-0.001,0.016)	1.8	0.132	0.133	-0.001 (-0.004,0.002)	-0.3	0.053	0.053
	Pooled logistic	0.007 (0.000,0.016)	2.3	0.131	0.132	0.017 (0.013,0.002)	5.6	0.055	0.058
0	Weighted Cox	0.005 (-0.004,0.014)	NA	0.144	0.144	0.002 (-0.001,0.005)	NA	0.053	0.053
	Pooled logistic	0.004 (-0.005,0.013)	NA	0.143	0.143	0.002 (-0.001,0.005)	NA	0.055	0.055
-0.3	Weighted Cox	0.002 (-0.008,0.012)	-0.8	0.161	0.161	0.001 (-0.002,0.004)	-0.4	0.054	0.054
	Pooled logistic	0.000 (-0.010,0.010)	0.1	0.160	0.160	-0.012 (-0.016,-0.009)	4.0	0.056	0.057

$L(j)$, underestimated the protective effect of current treatment by nearly 20% due to uncontrolled confounding by indication. Moreover, Model 2 yielded a seriously biased (41.5% relative under-estimation) estimate of the indirect effect of previous treatment.

As shown in Table 4, all MSMs yielded practically unbiased estimates for the current treatment effect $A(j)$ (relative bias $< 4\%$ in rows 3 to 7). For the indirect treatment effect $A(j-1)$, the marginal structural Cox models with the unstabilized weights (w_i) and the normalized unstabilized weights ($w_i^{(n)}$) gave strongly biased estimates (3rd and 4th rows), likely a consequence of the choice of parameters in the data-generating model as we discuss in Section 5 below. Notice that because the true effect of the previous treatment (HR = -0.112) is weak relative to the effect of current treatment (HR = -0.693), the differences in the absolute bias are less marked. The pooled logistic regression, with the stabilized weights (last row in Table 4) yielded much more biased estimates than the weighted Cox regression with the same type of weights (5th row), for which the 95% confidence intervals for the bias included 0. Normalization of the stabilized weights further reduced the bias from 11.2% to only 1% (6th row), yielding the most accurate estimate of the indirect effect of previous treatment.

The high variability of the MSM estimates is reflected in empirical standard deviations (SDs) and root mean squared errors (RMSEs) being systematically several times higher than corresponding SDs and RMSEs for both conventional (unweighted) models (Table 4). This variance inflation is particularly dramatic for the MSM that uses the unstabilized un-normalized weights in equation (2) while the lowest values of SD and RMSE are obtained in the IPTW Cox regressions that use normalized weights.

Interestingly, for both direct $A(j)$ and indirect $A(j-1)$ treatment effects, when using the stabilized weights, the weighted pooled logistic regression estimates show much higher variance and RMSE than the corresponding IPTW time-dependent Cox regression estimates (last row versus 5th row of Table 4).

Figure 1 further illustrates the differences between different models, comparing conventional Model 2 and three MSM models with respect to the differences between 2,000 point estimates and the true log hazard ratio for previous treatment $A(j-1)$. Conventional estimates are systematically biased but have dramatically lower variance than MSMs, while the MSM estimates are all free of marked bias (see Table 4 for exact values). Among MSMs, the unstabilized weights result in extremely unstable estimates, as expected. The stabilized weights reduce the variance considerably but yield a small number of outliers which are eliminated by normalization.

Table 4: Comparison of the performance of the different weighting schemes for estimating the marginal structural Cox PH model and pooled logistic regression with the stabilized weights using the HIV cohort simulation based on 2,000 samples, with 2,500 subjects and 10 observation times in each sample.

	$A(j)(\text{true: } -0.693)$				$A(j-1)(\text{true: } -0.112)$			
	Bias (95% CI)	Relative Bias(%)	SD	RMSE	Bias (95% CI)	Relative Bias(%)	SD	RMSE
Model 1 ¹	-0.002 (-0.006,0.003)	0.3	0.096	0.096	0.113 (0.109,0.117)	-100.7	0.099	0.150
Model 2 ²	0.134 (0.130,0.138)	-19.3	0.091	0.162	0.047 (0.042,0.051)	-41.5	0.101	0.111
MSM ³ (weight w)	-0.003 (-0.036,0.030)	0.4	0.755	0.755	0.126 (0.092,0.160)	-112.4	0.785	0.795
MSM ³ (weight $w^{(n)}$)	0.026 (-0.007,0.046)	-3.8	0.438	0.439	0.064 (0.043,0.086)	-57.5	0.484	0.488
MSM ³ (weight $w^{(s)}$)	-0.009 (-0.013,0.031)	1.4	0.504	0.504	-0.013 (-0.036,0.011)	11.2	0.543	0.543
MSM ³ (weight $w^{(ns)}$)	0.020 (-0.001,0.040)	-2.8	0.458	0.459	-0.001 (-0.023,0.020)	1.0	0.488	0.488
MSM (logistic) ⁴ (weight $w^{(s)}$)	-0.023 (-0.054,0.008)	3.3	0.711	0.712	0.042 (0.009,0.075)	-37.5	0.755	0.756

¹Model 1: adjusted for the current value of the confounder

²Model 2: adjusted for the baseline value of the confounder

³MSM: marginal structural Cox model estimated using weighted time-dependent Cox regression

⁴MSM (logistic): marginal structural Cox model estimated using weighted pooled logistic regression

Figure 1: Estimates of the effect of $A(j-1)$ from four of the models reported in Table 4 (white line is the true value, black line is the mean of estimates): (a) unweighted Cox model, (b) MSM with the unstabilized weights, (c) MSM with the stabilized weights, and (d) MSM with the normalized stabilized weights.

4.2.2 Results of sensitivity analyses

In sensitivity analyses, we varied the effects of current treatment or of the confounder on the hazard, as well as the sample size. Table 5 investigates the impact of the strength of the effect of current treatment on the bias, the standard deviation, and the RMSE for the two treatment effects, estimated using different models. In particular, it compares conventional Model 2, MSMs estimated with weighted time-dependent Cox regression using stabilized or normalized stabilized weights, and weighted pooled logistic regression with stabilized weights.

In (conventional) Model 2, as the current treatment effect becomes weaker, the relative underestimation bias for the effect of $A(j)$ increases proportionally, indicating that the absolute bias remains stable (three first rows of Table 5). In general, the comparison between different estimates is not markedly affected by the strength of the treatment effect: all MSMs eliminated the bias but inflated the variance of the estimates relative to Model 2. As in Table 4, the variance inflation is more serious in the weighted pooled logistic regression than in the weighted time-dependent Cox regression despite using the same (stabilized) weights; the smallest SD is exhibited by estimators from the Cox model using the normalized stabilized weights (Table 5).

Further sensitivity analyses indicated that assuming different effects of the confounder on the hazard and, thus, varying the strength of the indirect treatment effect, did not markedly affect the results either (data not shown). Finally, with increasing sample sizes, the SDs and the RMSEs for all models decreased, as expected, while the bias did not change systematically (data not shown).

5 Discussion

In this paper, we adapted the MSM methodology to fit a marginal structural Cox model directly via IPTW in standard statistical software. We proposed a novel weighting approach and evaluated the performance of the proposed IPTW time-dependent Cox regression estimation of the marginal structural Cox model in a series of simulation experiments.

In our preliminary simulations, which replicated the assumptions and the data generation procedures used in a recent simulation study by Young et al. (2009), the marginal structural Cox model estimated via IPTW in a Cox regression yielded somewhat more accurate estimates of the causal effect of the treatment than the pooled logistic regression approximation, especially

Table 5: Sensitivity analyses: The impact of the strength of the effect of current treatment in a conventional Cox model, a marginal structural Cox model estimated via IPTW in Cox regression with standardized or normalized weights, and a marginal structural Cox model estimated via IPTW in pooled logistic regression with the stabilized weights for the HIV cohort simulation based on 2,000 samples, with 2,500 subjects and 10 observation times in each sample.

Model	log(HR) of $A(j)$	$A(j)$				$A(j-1)$ (true: -0.112)			
		Bias (95% CI)	Relative Bias(%)	SD	RMSE	Bias (95% CI)	Relative Bias(%)	SD	RMSE
Model 2 ¹	-1.204	0.134	-11.1	0.101	0.168	0.045	-39.8	0.102	0.111
	-0.693	0.134	-19.3	0.091	0.162	0.047	-41.5	0.101	0.111
	-0.357	0.137	-38.3	0.088	0.162	0.049	-43.9	0.096	0.108
MSM ² (weight $w_i^{(s)}$)	-1.204	-0.010	0.9	0.556	0.556	-0.031	27.3	0.580	0.581
	-0.693	-0.009	-1.4	0.504	0.504	-0.013	11.2	0.543	0.543
	-0.357	0.009	-2.6	0.506	0.506	-0.018	15.6	0.533	0.533
MSM ² (weight $w_i^{(ns)}$)	-1.204	0.011	-0.9	0.509	0.509	-0.015	13.7	0.523	0.523
	-0.693	0.020	-2.8	0.458	0.459	-0.001	1.0	0.488	0.488
	-0.357	0.027	-7.6	0.446	0.482	0.010	-9.0	0.482	0.482
MSM (logistic) ³ (weight $w_i^{(s)}$)	-1.204	-0.038	3.2	0.780	0.781	0.008	-7.1	0.827	0.827
	-0.693	-0.023	3.3	0.711	0.712	0.042	-37.5	0.755	0.756
	-0.357	0.002	-0.5	0.751	0.751	0.013	-11.7	0.725	0.726

¹Model 2: adjusted for the baseline value of the confounder

²MSM: marginal structural Cox model estimated using weighted time-dependent Cox regression

³MSM (logistic): marginal structural Cox model estimated using weighted pooled logistic regression

when the outcomes were relatively frequent. The potential advantages of the weighted marginal structural Cox regression were confirmed in our main simulations, as well as in sensitivity analyses, where it considerably reduced the variance and the RMSE of the estimators of both direct and indirect treatment effects relative to pooled regression MSM model with the same (stabilized) weights.

Simulations allowed us also to assess the impact of the choice of IPTW weight on the accuracy of the estimators obtained from the marginal structural Cox regression. First, our results clearly demonstrate the need to use the stabilized weights, proposed by Robins et al. (2000), to avoid numerical instability of the MSM estimates obtained using the naïve unstabilized weights, which may assign extreme weights to individual subjects who received unusual treatment given their covariate values. In our simulations, the stabilized weights significantly reduced the empirical SDs of the MSM estimators by a factor close to one. Furthermore, in addition to using frequently used unstabilized and stabilized weights (Hernán et al., 2001), we proposed a simple normalization of both types of weights that ensured that the mean weight for each assessment time was equal to 1. This allowed us to avoid potentially very extreme (large) unstabilized weights over time. Another *a priori* rationale for considering the normalized weights was that we expected that, by reducing sampling variation of the estimated time-dependent weights, these weights might also increase the efficiency of the resulting (weighted) estimators of the treatment effects. The results of our main simulations seem to corroborate these expectations as, among different versions of the weighted time-dependent Cox regression, the model with the normalized stabilized weights yielded the lowest empirical SD and RMSE values for both direct and indirect treatment effects.

Our main simulations attempted to mimic the longitudinal study of HIV progression, with time-varying treatment decisions that depended strongly on a time-varying confounder. The treatment was assumed to affect the hazard in two ways, through: (i) a direct effect of the current treatment, and (ii) an indirect effect of previous treatment, mediated through the change in the time-dependent confounder. The results confirm that, in the presence of a time-varying confounder that is also affected by the treatment, the (unweighted) conventional Cox proportional hazards model is unable to accurately estimate the causal effect of time-varying treatment on the hazard. Specifically, the Cox model that adjusts only for the baseline value of such a confounding/mediating variable may yield seriously biased estimates of the effect of current treatment. This is a reflection of confounding by indication, whereby subjects with a worse prognosis are typically more likely to be treated. On the other hand, the con-

ventional model that does adjust for the current covariate value of a covariate, i.e. that accounts for its changes over time, yields an unbiased estimate of the direct effect of current treatment but may seriously underestimate the indirect effect of previous treatment if these are mediated through the changes in time-varying covariates. Our simulations also confirm the well-established advantages of the MSM approach in estimating causal effects of time-varying treatments (Hernán et al., 2000). In contrast to biased estimates obtained with conventional unweighted Cox models, in all the simulated scenarios, all the MSM models yielded unbiased estimates of the direct effects of current treatment. Moreover, the time-dependent Cox regression with the normalized stabilized weights also yielded unbiased estimates of the indirect effect of previous treatment.

However, in our main simulations, the MSM estimators of both direct and indirect effects of time-varying treatment had systematically much higher variance than the corresponding estimators obtained from conventional (unweighted) Cox models. This resulted in the unbiased but unstable MSM estimates having systematically higher RMSEs than the corresponding, biased conventional estimates. We note that similar bias-variance trade-offs can be found in many areas of statistical research, where more refined and thus more complex models often reduce or eliminate bias but induce serious variance inflation. For example, a similar pattern was found in a recent simulation study that evaluated the performance of the instrumental variable (IV) approach in controlling for unobserved confounding in pharmaco-epidemiology: the IV estimation removed bias but often yielded higher RMSE than biased conventional estimates (Ionescu-Ittu et al., 2009). Our findings regarding RMSE differ from the simulations reported by Bryan et al. (2004), who fitted pooled logistic models to time-to-event data. They found that the MSM estimator had better efficiency (lower RMSE) than the unweighted model. However, as the censoring rate increased, the RMSEs of the two estimators became closer (Bryan et al., 2004). In our main simulations, the censoring rate (60%) is relatively high, which is clinically plausible given the short follow-up (5 years) and this may be one of the sources of the inflated variance. Further studies on the impact of censoring rates on the performance of MSMs are needed.

The high variability of the MSM estimators could be partly explained by the strong association between the confounder $L(j)$ and the treatment $A(j)$ assumed in our simulations, which resulted, for example, in the estimated probability of a subject with $L(j) > 500$ receiving the treatment $A(j) = 1$ being as low as 0.6%. Likewise, subjects with very low values of $L(j)$ had very low estimated probabilities of not being treated. This resulted in extremely large values of the weights w_i , especially in the case of the unstabilized weights,

for some person-visits. The problem of extreme weights is exacerbated in a dataset with long follow-up, as with each consecutive visit, probabilities continue to be multiplied together. Those subjects with extreme weights will then dominate the weighted analysis, since the resulting pseudo-population will include a very large number of copies of those unusual subjects. Because such subjects, with highly unexpected values of the actual treatment, will represent very rare events, their occurrence, and the actual patterns of their treatment history and outcomes, will vary substantially among simulated samples, inducing large sample-to-sample variability of the estimates of treatment effects in the weighted models. The stabilized weights ($w_i^{(s)}$) use, in the numerator, the probability of the observed treatment estimated while conditioning on the baseline covariates and treatment history, though not on the time-dependent confounder. This avoids assigning very high weights to those subjects whose observed treatment is unexpected due to inconsistency with the baseline covariate vector or treatment history. As reported in Section 4.2, this reduces dramatically the variability of the stabilized weights relative to the unstabilized weights and thus improves considerably the efficiency of the estimates. The proposed normalization of the stabilized weights ($w_i^{(ns)}$) avoids the tendency for differential weighting of earlier versus later observations, thereby further reducing the sampling variance of the estimated weights, which resulted in the lowest variance and RMSE among the four marginal structural Cox models estimated via weighted Cox regression.

The treatment assignment mechanism in the HIV cohort simulation may border on a violation of the experimental treatment assignment (ETA) assumption, also called the positivity assumption (Wang et al., 2006), which requires a positive probability of receiving every level of exposure for every combination of values of the past exposure and covariate histories that occur among individuals in the population. Cole et al. (2008) demonstrate that violation of the ETA assumption will increase the bias and variance of the causal parameters estimators and a strong association between confounder and exposure will make the violation of this assumption more likely. We observed this in the simulations that used Young’s data-generating procedure when parameters were selected to create a strong dependence of treatment assignment on covariates. To further explore this issue, we performed a post hoc sensitivity analysis, in which we assumed a weaker association between the confounder $L(j)$ and the treatment $A(j)$ (e.g., in the new simulation, $P[A = 1|L \geq 500]$ was about 30%, at the first visit, as opposed to $< 1\%$ in the original simulations). As expected, this reduced the variation of the MSM estimates: their SDs were now less than two times higher than SDs of the corresponding conventional estimates (data not show), as opposed to the SD ratios of 4.5 to 5

observed in our main simulations, reported in Table 4.

Some limitations of our study have to be recognized. As all simulation studies, we relied on some simplifying assumptions: while we attempted to mimic general features of a longitudinal study of HIV progression, the assumed causal structure of our data was relatively uncomplicated. Specifically, we assumed that the decision whether to treat a patient at a given visit depended only on the current value of the time-varying covariate and the treatment assigned at the previous visit. Furthermore, we assumed that the hazard in each interval between two visits remained constant and depended only on the most recent values of the treatment and the time-varying covariate, measured at the beginning of the interval. In practice, both the treatment decision and the hazard are likely to also depend on cumulative effects of past treatments, past history of changes in the time-varying covariate, their response to past treatments, and other covariates. Our assumptions implied considerable within-individual visit-to-visit variation in the current treatment status $A_i(j)$. Such variation may be more plausible for treating short-term medical conditions such as intermittent episodes of insomnia or anxiety treated with psychotropic medications. Indeed, psychotropic drugs are usually prescribed for less than three months and individual subjects have often several periods of use and non-use (Bartlett et al., 2004). In contrast, in most applications of MSMs to study HIV progression, it is assumed that the patient's treatment status may change only once, i.e. once the treatment was initiated, the patient is treated until the end of his/her follow-up (Hernán et al., 2000; Cole et al., 2003; Sterne et al., 2005; Bryan et al., 2004). However, because of partial non-compliance, permanent treatment discontinuation, and/or physician's concerns about the risk of developing drug resistance, in reality, the actual treatment sequences may fall somewhere between our model with large temporal variation and the one-change (step function) pattern assumed in previous publications. Further studies are necessary to investigate to what extent the systematic variance inflation of the MSM estimates observed in our simulations may be due to the large variation over time within individual $A_i(j)$ vectors and resulting within-subject variation in time-dependent weights.

Another crucial arbitrary assumption underlying our main simulations implied that the total effect of treatment on the logarithm of the hazard may be decomposed into two additive components: direct effect of current treatment and indirect effect of the treatment at the previous visit, entirely mediated through the change in the time-dependent covariate. This critical assumption largely facilitated generation of survival time conditional on the current values of the time-varying covariate and treatment, and assessment of the accuracy of the estimates. Future studies should consider more complex models for the

causal effect of the treatment. A recently developed permutational algorithm for generating event times conditional on arbitrarily complex time-dependent covariates and/or effects (Sylvestre and Abrahamowicz, 2008) may be useful to simulate such data structures (Burton et al., 2006).

Our (limited) sensitivity analyses suggested that our results and conclusions are robust with respect to sample size, and treatment or confounder effects on the hazard. Future simulations should investigate wider ranges of these parameters, as well as the potential impact of the number of visits on the results. Further refinements of our estimation methods should also be considered. For example, it will be interesting to assess whether truncating extreme values of the estimated weights could reduce the variance of the MSM estimates of the treatment effects while keeping these estimates free of bias. Cole et al. (2008) report encouraging empirical results obtained with truncated weights.

In conclusion, our simulations confirm the ability of the marginal structural Cox models to remove bias due to time-dependent covariates that act as both confounders and mediators of the causal treatment effects. The marginal structural Cox model, implemented using weighted time-dependent Cox regression, yielded lower standard deviations of the treatment effect estimators than the pooled logistic regression MSM approximation, less biased estimates in scenarios with more frequent events, and more accurate estimates for the indirect treatment effect. Our results demonstrate the importance of using the stabilized weights to both eliminate the bias and reduce the variance of the estimators, and suggest that the proposed normalization of the stabilized time-dependent weights may further reduce the variance. Nevertheless, in most simulated scenarios, the variance of the MSM estimators was seriously inflated relative to conventional (unweighted) model estimators. Future research in this area should focus on new methods to further reduce the variance of the MSM estimators.

References

- G. Bartlett, M. Abrahamowicz, R. Tamblyn, R. Grad, R. Capek, and R. du Berger. Longitudinal patterns of new Benzodiazepine use in the elderly. *Pharmacoepidemiology and Drug Safety*, 13(10):669–682, 2004.
- J. Bryan, Z. Yu, and M.J. van der Laan. Analysis of longitudinal marginal structural models. *Biostatistics*, 5(3):361–380, 2004.

- A. Burton, D.G. Altman, P. Royston, and R.L. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292, 2006.
- H.K. Choi, M.A. Hernán, J.D. Seeger, J.M. Robins, and F. Wolfe. Methotrexate and mortality in patients with rheumatoid arthritis: a prospective study. *The Lancet*, 359(9313):1173–1177, 2002.
- S.R. Cole and M.A. Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664, 2008.
- S.R. Cole, M.A. Hernán, J.M. Robins, K. Anastos, J. Chmiel, R. Detels, C. Ervin, J. Feldman, R. Greenblatt, L. Kingsley, et al. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology*, 158(7):687–694, 2003.
- N.R. Cook, S.R. Cole, and C.H. Hennekens. Use of a marginal structural model to determine the effect of aspirin on cardiovascular mortality in the Physicians’ Health Study. *American Journal of Epidemiology*, 155(11):1045–1053, 2002.
- M.H. Gail. Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In Moolgavkar S.H. and Prentice R. L., editors, *Modern Statistical Methods in Chronic Disease Epidemiology*, pages 3–18, New York, 1986. Wiley.
- N.M. Graham, S.L. Zeger, L.P. Park, S.H. Vermund, R. Detels, C.R. Rinaldo, and J.P. Phair. The effects on survival of early treatment of human immunodeficiency virus infection. *New England Journal of Medicine*, 326(16):1037–1042, 1992.
- S. Greenland, J.M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46, 1999.
- M.A. Hernán, B. Brumback, and J.M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570, 2000.
- M.A. Hernán, B. Brumback, and J.M. Robins. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96(454):440–448, 2001.

- R. Ionescu-Ittu, J.A. Delaney, and M. Abrahamowicz. Bias-variance trade-off in pharmacoepidemiological studies using physician-preference-based instrumental variables: a simulation study. *Pharmacoepidemiology and Drug Safety*, 18(7):562–571, 2009.
- S. Kinloch-de Loes, B.J. Hirschel, B. Hoen, D.A. Cooper, B. Tindall, A. Carr, J.H. Saurat, N. Clumeck, A. Lazzarin, L. Mathiesen, et al. A controlled trial of zidovudine in primary human immunodeficiency virus infection. *The New England Journal of Medicine*, 333(7):408–413, 1995.
- J. Ritz and D. Spiegelman. Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research*, 13(4):309–323, 2004.
- J.M. Robins. Marginal structural models. 1997 Proceedings of the section on Bayesian statistical science. *American Statistical Association*, pages 1–10, 1997.
- J.M. Robins. Association, causation, and marginal structural models. *Synthese*, 121(1):151–179, 1999a.
- J.M. Robins. Marginal structural models versus structural nested models as tools for causal inference. *Statistical Models in Epidemiology: The Environment and Clinical Trials*, 116:95–134, 1999b.
- J.M. Robins and M.A. Hernán. Estimation of the causal effects of time-varying exposures. In Fitzmaurice G, Davidian M, Verbeke G, and Molenberghs G, editors, *Longitudinal Data Analysis*, pages 553–599. CRC Press, 2008.
- J.M. Robins, M.A. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.
- J.A.C. Sterne, M.A. Hernán, B. Ledergerber, K. Tilling, R. Weber, P. Sendi, M. Rickenbach, J.M. Robins, and M. Egger. Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. *The Lancet*, 366(9483):378–384, 2005.
- M.P. Sylvestre and M. Abrahamowicz. Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in Medicine*, 27(14):2618–2634, 2008.
- R Development Core Team. R v2.8.0. *R Foundation for Statistical Computing, Vienna, Austria.*, 2008. URL <http://www.R-project.org>.

- Y. Wang, M.L. Petersen, and D. Bangsberg. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. 2006. URL
- D. Westreich, P. MacPhail, A. Van Rie, B. Malope-Kgokong, P. Ive, D. Rubel, R. Boulmé, J. Eron, and I. Sanne. Effect of pulmonary tuberculosis on mortality in patients receiving HAART. *AIDS*, 23(6):707–715, 2009.
- J.G. Young, M.A. Hernán, S. Picciotto1, and J.M. Robins. Relation between three classes of structural models for the effect of a time-varying exposure on survival. *Lifetime Data Analysis*, 16(1):71–84, 2009.