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

Comparison of Approaches to Weight Truncation for Marginal Structural Cox Models

Abstract: Marginal structural Cox Models (Cox MSMs) have been used to estimate the causal effect of a time-varying treatment on the hazard when there exist time-dependent confounders, which are themselves also affected by previous treatment. A Cox MSM can be estimated via the inverse-probability-of-treatment weighting (IPTW) estimator. However, IPTW estimators suffer from large variability if some observations are assigned extremely high weights. Weight truncation has been proposed as one simple solution to this problem, but truncation levels are typically chosen based on ad hoc criteria that have not been systematically evaluated. Bembom et al. proposed data-adaptive selection of the optimal truncation level using the estimated mean-squared error (MSE) of a truncated IPTW estimator for cross-sectional data. Based on a similar principle, we proposed data-adaptive approaches to select the truncation level that minimizes the expected MSE for time-to-event data with time-varying treatments. The expected MSE is approximated by using either observed statistics as a proxy for the true unknown parameter or using cross-validation. Simulations confirm that simple weight truncation at high percentiles such as the 99th or 99.5th of the distribution of weights improves the IPTW estimators in most scenarios we considered. Our newly proposed approaches exhibit similarly good performance and may be applied in a wide range of settings.

Keywords: marginal structural model, survival analysis, IPTW estimator, positivity assumption, weight truncation

***Corresponding author: Erica E.M. Moodie**, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada, E-mail: erica.moodie@mcgill.ca
Yongling Xiao, The Institut national d'excellence en santé et en services sociaux (INESSS), Montreal, Canada, E-mail: yongling.xiao@.mcgill.ca
Michal Abrahamowicz, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada, E-mail: michal.abrahamowicz@mcgill.ca

1 Introduction

In survival analysis, time-dependent predictors of survival that are also predictors of subsequent treatment act as time-dependent confounders. Some time-dependent confounders are also mediating variables, i.e., treatment modifies the subsequent values of such a variable, which in turn changes the risk of the outcome. If one then wishes to estimate the direct and indirect effects of a time-varying sequence of treatments on survival, the causal structure of the data implies that the analysis should not adjust for such a variable, since adjusting for a mediating variable would bias the estimation of the indirect (and total) treatment effect (Robins et al., 2000). However, if the analysis does not adjust for this variable, the estimator of the direct treatment effect will be biased due to confounding by indication. Thus, the conventional multivariable Cox proportional hazards (PH) model (Cox, 1972) cannot yield an unbiased estimate of the total causal effect of treatment whenever time-varying confounders act also as mediating variables.

Marginal structural models (MSMs), a class of models, that are used to model the marginal distribution of counterfactual variables, have been proposed to define the causal effect of a time-varying treatment in such challenging situations (Robins, 1997, 1999a,b; Robins et al., 2000; Robins and Hernán, 2008). For time-to-event data, a marginal structural Cox PH model (Cox MSM) can be applied and consistently estimated via the inverse-probability-of-treatment weighting (IPTW) estimator. Using the IPTW approach, time-dependent confounding is controlled by assigning a weight to each subject or observation which is inversely proportional to his probability of receiving the observed treatment given the past history of treatment and confounders as well as the baseline covariates (Hernán et al., 2000, 2001).

However, the consistency of the IPTW estimator relies heavily on the positivity assumption (also referred to as the experimental treatment assignment (ETA) assumption), which states that each subject in the study population has a positive probability of being exposed to each level of the treatment, regardless of his history of the past treatment and the past vectors of all the relevant covariates (Neugebauer and van der Laan, 2005; Wang et al., 2006; Cole and Hernán, 2008). This assumption may be *theoretically* or *practically* violated. When all subjects with a given level of covariates are certain to receive a particular treatment, positivity is theoretically violated and the IPTW estimator will be inconsistent (Robins et al., 2000; Cole and Hernán, 2008; Neugebauer and van der Laan, 2005). Positivity is practically violated when the probability that subjects in a subgroup corresponding to a particular combination of covariates receives a particular treatment is extremely low, so that only very few such subjects in a given study sample are observed to receive this treatment (Cole and Hernán, 2008). In the latter situation, those subjects with a very low probability of receiving the treatment that they did in fact receive will be assigned extreme large weights. Practical ETA violations result in highly influential observations, leading to the instability of the IPTW estimator (Wang et al., 2006). Practical ETA violations are of particular concern in studies with long follow-up times, as the weights may be magnified by the multiplication of several very small probabilities.

Stabilized weights and the additional normalization of stabilized weights over the follow-up time can be used to reduce the variability of the weights (Robins et al., 2000; Xiao et al., 2010). However, when there are very strong covariate-treatment associations, the IPTW estimator based on the stabilized weights will still be highly variable and will have a positively skewed distribution (Hernán et al., 2000; Neugebauer and van der Laan, 2005). Several other approaches have been proposed to deal with ETA violation. These include (i) truncation, whereby all the weights larger than the user-specified maximum weight are replaced by that threshold value (Kish, 1992; Wang et al., 2006; Bembom and van der Laan, 2008b; Cole and Hernán, 2008; Moore et al., 2009), (ii) use of non-IPTW estimators, such as G-computation (van der Wal et al., 2009), (iii) exclusion of covariates that cause serious weight inflation but are only weakly associated with the outcome (Bembom and van der Laan, 2008a), (iv) exclusion of those observations that violate the ETA assumption (sometimes referred to as “trimming”) (LaLonde, 1986; Heckman et al., 1998; Dehejia and Wahba, 1999; Crump et al., 2006), and (v) history-restricted MSMs (HR-MSM), that estimates the causal effect of the treatment on the outcome based on a curtailed portion of the treatment history rather than the entire treatment history (Neugebauer et al., 2007).

In this paper, we investigate approaches to weight truncation for a Cox MSM when ETA assumption is practically violated. We assess the performance of several existing easy-to-implement methods (Cole and Hernán, 2008), as well as some novel approaches that we proposed to select an optimal weight truncation level. The paper is organized as follows: In Section 2, we first review the marginal structural Cox model, and then we describe alternative weight truncation approaches. In Section 3, we describe the design and methods of our simulation study. Section 4 summarizes the results of the simulations, comparing the performance of the proposed weight truncation methods with an untruncated Cox MSM, and the fixed-value or fixed-percentile weight truncation approaches as well as the newly proposed data-adaptive choice of optimal weight truncation. In Section 5, we apply the various truncation methods to the Multicenter AIDS Cohort Study (MACS) data, to assess the effect of treatment with Zidovudine (AZT) on AIDS-related deaths. We conclude with a discussion of the results (Section 6), the limitations of our study, and suggestions for future research.

2 The marginal structural Cox model

2.1 Notation and definition

We consider a longitudinal study in which N subjects are followed at regular time intervals (e.g., every 6 months) from entry into the cohort (time zero), and at each follow-up visit k, $k = 1,\ldots, m$, subjects' covariates are measured and their current treatment assignments are then decided. We denote by T_i, the observed follow-up time of subject i, and V_i, the vector of baseline covariates, measured at time zero. For subject i at time t, let $A_i(t)$ be a binary indicator of the received treatment and $L_i(t)$ be the current value of the time-dependent covariate L. Since subjects are assessed only at discrete time intervals, we assume that the treatment $A_i(t)$ and covariate $L_i(t)$ remain the same within the interval between any two adjacent assessments $(k, k+1)$. Let $Y_i(t) = 1$ if subject i had an event at time t, and 0 otherwise. Those who are either lost to follow-up or do not have an event until the end of study are rightly censored at the respective time. We use an overbar to represent a covariate history, thus, $\bar{A}(t)$ indicates the treatment history up to t, and $\bar{L}(t)$ indicates the history of the time-dependent confounder until that time.

Let $T_{\bar a}$ denote a random variable representing a subject's counterfactual survival time, followed a given treatment history $\bar a(t) = \{a(u);0 \le u \le t\}$. For each $\bar a$, the COX MSM is defined as (Hernán et al., 2000) follows:

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

where $\lambda_{T_{\bar a}}(t|V)$ is the hazard at time t for a subject with the baseline covariate vector V had, contrary to fact, the subject followed the treatment history $\bar a$; β_1 and β_2 are unknown causal parameters; $\lambda_0(t)$ is the unspecified baseline hazard at time t for a subject who is never treated $\bar a(t) = \bar 0$, with $V = 0$; and $f(\cdot)$ is an analyst-defined function of the treatment history. The choice of baseline covariates to include in an MSM falls to the analyst. Of course, because the true form of the model is not known, as the number of included covariates increases, the risk that some of their effects are misspecified increases as well.

As noted above, the causal parameters β of a MSM can be estimated using IPTW to account for the time-dependent confounding effect of $\bar L(t)$ (Robins, 1997, 1999a). A commonly used weight is the *stabilized* weight, $w_i^{(s)}(t)$, defined by Robins et al. (2000), as follows:

$$w_i^{(s)}(t) = \prod_{k=1}^{m(t)} \frac{P\left[A(k) = a_i(k)|\bar A(k-1) = \bar a_i(k-1), V = v_i\right]}{P\left[A(k) = a_i(k)|\bar A(k-1) = \bar a_i(k-1), \bar L(k) = \bar l_i(k), V = v_i\right]}, \qquad [2]$$

where $m(t)$ is the total number of visits up to t, including the initial visit ($k = 1$, $t = 0$). The denominator of the stabilized weight is the probability that a subject received his own observed treatment at time t, $A(t)$, given his own past treatment history, confounder history, and baseline covariates. The numerator, the probability that a subject received his observed treatment at time t conditional on only his past treatment history and baseline covariates, but not on time-dependent confounder $\bar L(t)$, is used to reduce the variability of the original "unstabilized"weights (Hernán et al., 2000). By accounting the 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 treatment.

Four assumptions are needed to consistently estimate the causal parameters with the IPTW estimator: consistency, exchangeability, no-model misspecification, and positivity (Cole and Hernán, 2008). The consistency assumption is the fundamental assumption that links the counterfactual data $T_{\bar a}$ to the observed data $T_{\bar A}$; it states that the observed outcome $T_{\bar A}$ is equal to the counterfactual outcome $T_{\bar a}$ under the observed treatment history, that is, $\bar a = \bar{\mathrm{A}}$. Exchangeability states that given the recorded covariates in V and $L(t)$, there are no other unmeasured confounders. The no-model misspecification assumption postulates that the unknown probabilities $P\left[A(k) = a(k)|\bar A(k-1), \bar L, V\right]$ are modeled through a correctly specified model, and the marginal structural model is correctly specified. However, in practice, any unsaturated models may be misspecified, and thus may lead to biased causal effect estimates (Robins and Hernán, 2008). Neugebauer and van der Laan (2007) developed a nonparametric MSM (NPMSM) approach, which does not require correct specification of a parametric model but relies on a (potentially misspecified) working model instead.

2.2 Impact of the ETA violation on the IPTW estimates

A practical ETA violation can lead to some observations being assigned very large weights. To illustrate how observations with the most extreme weights (i.e., the most influential observations) can affect the IPTW estimates, we purposely selected four simulated samples that yielded IPTW estimates for the current treatment effect $A(j)$ that were exhibited a relative difference of more than 50% from the true parameter value. The highest weight in each selected sample was at least 40% higher than the second highest weight in the same sample.

For each selected sample, we generated $B = 100$ resamples, and for each resample, the untruncated IPTW estimate was calculated. The impact of the most highly weighted observation on the IPTW estimate can then be observed from the distribution of the resample-specific estimates by tracing whether a resample included this observation. In both top panels of Figure 1, it is evident that the resample-specific estimates have a strictly bimodal distribution: in all those resamples in which the observation with the extreme weight was included (dark dots) the treatment effect is highly overestimated, while all other resamples (light dots) yield estimates reasonably close to the true effect (indicated by the dashed line). In both samples, the event associated with the highest weight occurs for subjects who are treated and have high values of the confounder, which could suggest that treatment is associated with a higher hazard of the event. Thus, because of the extremely high weight assigned to these unusual observations, the weighted maximum partial likelihood estimator (MPLE) of the treatment effect yields a counter-intuitive log(HR) > 0. The impact of those observations on the IPTW estimates can also be illustrated by the dramatic change in the estimates that was observed when we just artificially reversed the survival status of the observation associated with highest weight from "event" to "no event": the estimated effects of current treatment $A(j)$ changed from 0.444 to −0.654 and from 0.665 to −1.179 for samples (a) and (b), respectively.

In contrast, the two bottom panels of Figure 1 show results for the two samples where those observations that are assigned extreme weights are *not* associated with the event. In those two samples, the inclusion or the exclusion of the observation with the highest weight has only a minor impact on the untruncated IPTW estimates. This confirms that censored individuals are less influential in the survival analysis.

The examples presented here indicate that a single observation may determine and indeed bias the results for the entire sample. This can be avoided by truncation of the weights, which reduces the influence of some observations, e.g., with fixed 99.5th percentile weight truncation, the estimated effects of current treatment $A(j)$ changed from 0.444 to −0.569 and 0.665 to −0.578 for samples (a) and (b), respectively. We are thus motivated to find a principled and optimal means of performing the truncation.

2.3 Alternative truncated IPTW estimators of the Cox MSM

2.3.1 The problem

Since the variability of the IPTW estimators is typically due to the impact of a few observations with above-illustrated extreme weights, weight truncation has been proposed to limit the maximum contribution that any observation in the data can have on the fitted Cox MSM at any time t (Cole and Hernán, 2008). Specifically, the truncated IPTW estimator relies on weights that are truncated at a prespecified constant M: $w_M^{(S)} = \min\left(w^{(s)}, M\right)$, i.e., any weight greater than M is replaced by the value (truncated weight) M (Cole and Hernán, 2008).

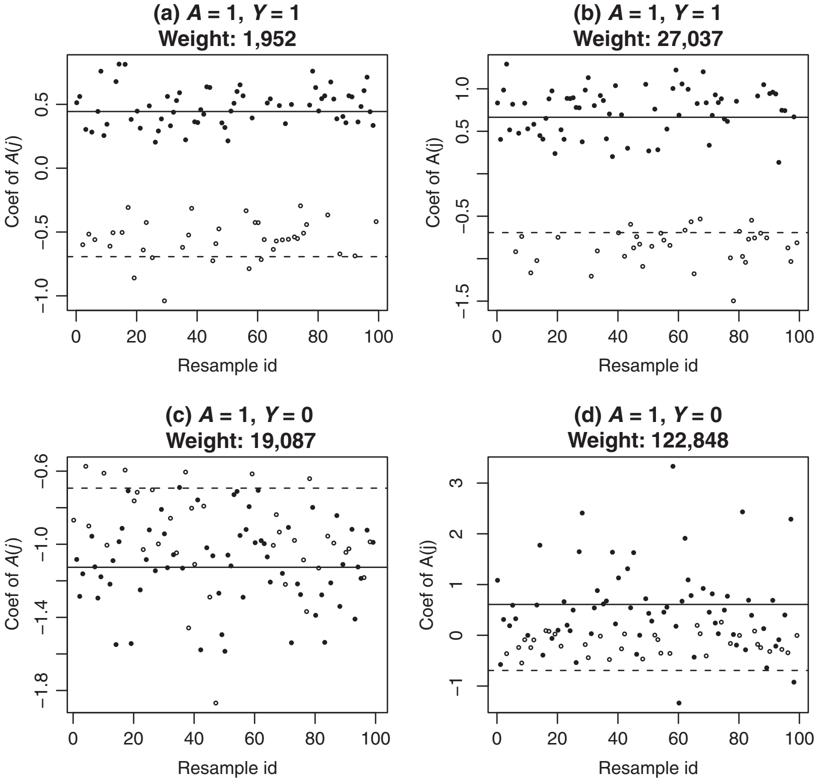

Figure 1 Investigation of the impact of the extreme weights on the IPTW estimates of *A*(*j*) using resampling. Dark dots indicate that the subject with the highest weight in the original sample was included in a resample. The solid line represents the estimated effect of *A*(*j*) in the original sample and the dashed line represents the true effect. The treatment (*A*), the event (*Y*), and the weight for the observation with the highest weight are indicated in the title of each panel.

The direct effect of weight truncation is the reduction in the instability of the IPTW estimator (Moore et al., 2009). However, a consequence of weight truncation is the introduction of bias in the estimated weights, which in turn leads to bias of the IPTW estimator of the treatment effect. Thus, weight truncation requires a bias-variance trade-off (Cole and Hernán, 2008). Ideally, an optimal truncation level should be selected so that the corresponding truncated IPTW estimator yields the minimum mean-squared error (MSE) (Bembom and van der Laan, 2008b). However, it is difficult, in practical applications, to assess the MSE of an IPTW estimator. Bembom and van der Laan (2008b) proposed a closed-form estimate for the expected MSE of a truncated IPTW estimator in order to select an appropriate truncation level; however, the approach is limited to cross-sectional data with either continuous or binary outcomes. Although it is possible to extend the approach for longitudinal data, the computation can become burdensome as the number of time intervals increases and the development of a similar closed-form equation for longitudinal, right censored time-to-event data is not straightforward.

2.3.2 Proposed empirical criteria to select the truncation level

Based on the principle of MSE minimization considered by Bembom and van der Laan (2008b), we propose two easily implementable approaches to approximate the expected MSE, based on which the optimal truncation level can be selected. Specifically, we estimate the expected MSE of the truncated IPTW estimator for each of several truncation levels considered, then select the truncation level that corresponds to the smallest MSE. The MSE of an estimator $\hat{\beta}_M$ is defined as follows:

$$\widehat{MSE}\left(\hat{\beta}_M\right) = \mathrm{SD}\left(\hat{\beta}_M\right)^2 + \left(\hat{\beta}_M - \beta\right)^2, \qquad [3]$$

where M denotes the truncation level, β denotes the true causal treatment effect, $\hat{\beta}_M$ denotes the truncated IPTW estimate, SD denotes the standard deviation of the estimator, and MSE denotes the mean-squared error.

In practice, it is impossible to calculate the true MSE because β is unknown. However, if one were able to find an estimable quantity which is monotonically associated with the true MSE, it would still be possible to identify the optimal truncation level based on this approximating quantity. We consider two approaches to select the optimal truncation level that approximate the true MSE (specifically, the bias component) in different ways. In both approaches, the variance component of MSE in eq. [3] is estimated by the square of the estimated standard error (SE) of the corresponding truncated IPTW estimator.

The first approach to approximate the bias component in eq. [3] uses an estimable "proxy" for the unknown β . The expected MSE is then calculated as follows:

$$\widehat{MSE}\left(\hat{\beta}_M\right) = \mathrm{SE}\left(\hat{\beta}_M\right)^2 + \left(\hat{\beta}_M - \tilde{\beta}\right)^2, \qquad [4]$$

where SE represents the standard error of the truncated ITPW estimator $\hat{\beta}_M$, and $\tilde{\beta}$ denotes the proxy, or substitute, for the true causal parameter β . Three candidates are considered for $\tilde{\beta}$: (i) the untruncated IPTW estimate; (ii) the truncated IPTW estimate with weights truncated at the 99.5th percentile of sample weights; and (iii) the truncated IPTW estimate based on the 99th percentile. The percentiles considered (99.5th and 99th) are consistent with Cole and Hernán (2008), who suggested truncating weights at the first and 99th percentiles. In preliminary simulations, no benefit was observed with truncation of weights at the left tail of the weight distribution. Thus in all the percentile-related approaches considered in this article, weights are truncated only in the right tail (that is, restricting only the maximum weights).

We also consider a second approach, in which the data are divided into halves, and the untruncated IPTW estimate from one subset is used as the proxy $\tilde{\beta}$ for the other half, and *vice versa*. That is, the data are first randomly divided into two equal-sized subsets. The MSE is then estimated using eq. [4] separately for each subset by using the untruncated IPTW estimate from the other subset as the proxy $\tilde{\beta}$. The average of the two estimated MSEs from the two subsets is used to approximate the expected MSE of the entire dataset. Since we require a large sample size to ensure a better estimate of the "truth" from one subset, increasing the number of folds may be risky as it may result in small sample size in each subset. Therefore, in order to improve the estimation, we considered repeating this two-fold procedure r times, taking the mean of the r MSEs as the final estimate of MSE. The rationale behind the repetition of the cross-validation (CV) $r \geq 2$ times is to reduce the possible impact of sampling error of the dissimilarity of the two halves of the datasets. We refer to this cross-validation-like method as the CV approach.

3 Simulation study: design and analysis

3.1 The data

We simulated a prospective study of a hypothetical cohort of N HIV-positive patients. Using the notation of Section 2.1, for the *ith* ($i = 1,\ldots, N$) patient, the event time, from the start of the follow-up to an AIDS-defining event, is denoted by T_i, measured in years. Patients were evaluated every 6 months, and at each visit, the decision to initiate, continue, or interrupt treatment was made. At each visit j, ($j = 1,\ldots, m$), the time-varying binary treatment, highly active antiretroviral therapy (HAART), is denoted by $A_i(j)$ where $A_i(j) = 1$ if treatment was received and $A_i(j)$=0 otherwise, and a time-dependent continuous confounder, CD4 cell count, is denoted by $L_i(j)$. Whether a subject had an event in the interval between visits $[j, j + 1)$ is denoted by $Y_i(j)$ with 1 indicating that the event occurred and 0 otherwise. The only baseline covariate was the pretreatment value of the confounder: $L_i(1)$. The time-dependent covariate $L_i(j)$ was assumed to act as both a confounder and a mediator for the treatment effect on survival throughout. The causal diagram is shown in Figure 2 with the subscript for omitted subject.

The data were generated using the same methods and data-generating distributions as used in Xiao et al. (2010). The data generation algorithm was briefly described in Appendix A.1, and the details are described in section 3.2.1 of Xiao et al. (2010). R code is available on request from the corresponding author. In brief, we generated baseline covariates from a uniform distribution, and then alternated between (i) generating visit-specific treatment decisions (on or off treatment) using a logistic model, which depended on the most recent confounder and treatment values and (ii) generating the time-varying confounder whose value depended on its most recent values and on the most recent treatment assignment. The expected time-to-event was generated using a conditional Cox PH model in which a subject's hazard of having an event at a given time depended only on his current values of confounders and treatment. The noninformative censoring time was generated from a uniform distribution. The observed time-to-event and the corresponding censoring status were then determined by comparing the expected time-to-event, the generated censoring time, and the maximum follow-up time. Note that, unlike the algorithm of Young et al. (2009), this data generation approach does not use the exact marginal causal parameter β . However, in situations such as the simulations we performed, where the event rate in any interval is small (on an average, about 4% in each of the ten between-visits intervals), the noncollapsibility of the hazard ratios is negligible and hence so is the difference between the marginal and conditional parameters.

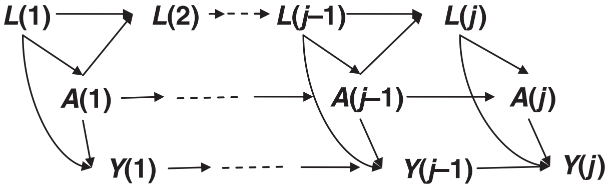

Figure 2 Causal diagram for the data-generating mechanism of the hypothetical cohort of HIV-positive patients used in the simulation study.

Two thousand independent random samples were generated, and sample sizes of $N = 400$, 2,000, and 10,000 were considered. The maximum follow-up time was 5 years ($m = 10$ visits). The average censoring rate for the main scenario, with $N = 2{,}000$, was approximately 60%. Appendix A provides information on the true models used to generate the data, the underlying assumptions, as well as the basic descriptive statistics on the distribution of the estimated inverse-probability-of-treatment (IPT) weights at selected time points.

3.2 Analysis of the simulated data

To estimate the total causal effect of current treatment $A(j)$ (which acts directly on the outcome) and past treatment $A(j - 1)$ (which acts indirectly via $L(j)$), on the time-to-event, we fitted the following marginal structural Cox model for each sample:

$$\lambda_{i,j}\left[t|\overline{A}_i(j), L_i(1)\right] = \lambda_0(t)\exp[\beta_1 L_i(1) + \beta_2 A_i(j) + \beta_3 A_i(j-1)] \qquad [5]$$

with weights truncated according to the different approaches described in Section 2.3.2.

We analyzed each simulated data using several different IPTW estimators of the Cox MSM, as given in Table 1. The untruncated IPTW estimator of the Cox MSM (Estimator 1) was compared with different weight truncation approaches, including weights truncated at (i) fixed values, (ii) fixed percentiles of the weight distribution in a given sample, and (iii) truncation levels selected using the proposed approaches described in Section 2.3.2. For fixed values of truncation, we considered a total of 23 values (spread between

Table 1 Description of different estimators of the marginal structural Cox model.

Description	Specification	Estimator	Label
Untruncated weights		1	
Weight truncated at a fixed percentile of weights distribution	99.5th	2	
	99.0th	3	
Weight truncated at the selected level giving minimum expected MSE, which was calculated with $\widetilde{\beta}$ =	Estimator 1	4	Untruncated
	Estimator 2	5	Perc995
	Estimator 3	6	Perc99
	2-fold cross-validation	7	CV
	True causal parameters	8	True

3 and 100: 3–15 incremented by 1, 20–50 incremented by 5, 75, and 100); for fixed percentiles, we used the 99.5th (Estimator 2) and 99th (Estimator 3) percentiles of the sample distribution of stabilized weights. Estimator 4 used the untruncated IPTW estimate as the proxy $\widetilde{\beta}$, while Estimators 5 and 6 used the 99.5th and 99th percentile truncated IPTW estimates as the proxy, respectively. Estimator 7 is based on the two-fold cross-validation approach with $r = 1$ or 5, and Estimator 8 corresponds to the "oracle scenario" that uses the true causal parameters in the calculation of the MSE. For each estimator, the mean of the MSEs for $A(j)$ and $A(j-1)$ was calculated to estimate the overall MSE for a given truncated estimator, and the optimal truncation level for a given sample was selected to correspond to the minimum estimated MSE.

Stabilized weights were used throughout and calculated for each person-visit, using eq. [2]. The denominator (i.e., the treatment assignment probability) was estimated with the correctly specified logistic regression model. Specifically, it was calculated as the estimated probability of receiving the observed treatment, $A(j)$, for subject i at visit j, conditional on the current value of the confounder, $L_i(j)$, and the indicator of the previous treatment, $A_i(j-1)$. The numerator was calculated as the probability of receiving the observed treatment estimated as a function of the baseline value, $L_i(1)$, and treatment history, $A_i(j-1)$, only. The marginal structural Cox models were then estimated via a time-dependent weighted Cox regression model adapted for the person-visits data (Xiao et al., 2010). The robust variance estimators were used to account for within-subject correlation induced by the use of time-dependent weights. All analyses were carried out using **R** (R Development Core Team, 2011).

The bias of the IPTW estimators was estimated as the mean difference between the 2,000 estimates and the corresponding true value of the respective causal parameter. The ratio of the SD of the 2,000 estimates to the mean SE was calculated to evaluate the accuracy of the SE estimators. To quantify the bias-variance trade-off, the root mean-squared error (RMSE) was calculated as the square root of the sum of squared bias and variance. The 95% coverage rate was calculated as the proportion of samples in which the nominal 95% confidence intervals (CIs) included the corresponding true parameter.

We assessed the performance of the proposed weight truncation schemes across different sample sizes (N = 400, 2,000, and 10,000). We also evaluated the performance of the proposed methods under different degrees of violation of the ETA assumption, which was controlled by modifying the association between the treatment and the confounder. A weaker treatment-confounder association induces a lesser ETA assumption violation, while a stronger one's association induces a greater violation.

4 Results

4.1 Main results

Figure 3 displays simulation results for Cox MSM parameters estimated with the stabilized weights truncated at alternative fixed values. The bias and the empirical SD over 2,000 estimates are reported. As expected,

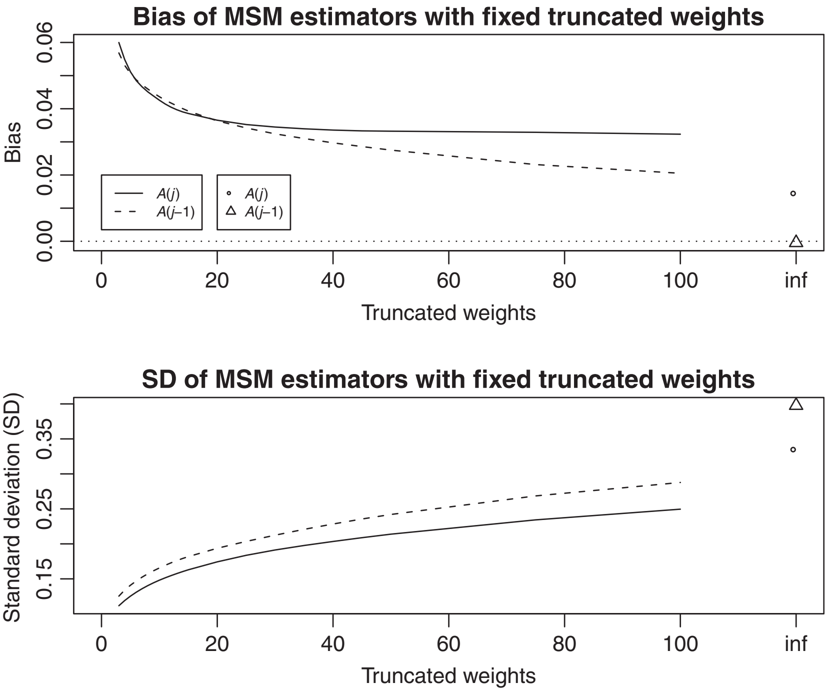

Figure 3 Properties of IPTW estimators of Cox MSM with the stabilized weight truncated at fixed values or not at all (“inf”).

the estimators of the Cox MSM that uses untruncated IPTW are unbiased; this corresponds to truncation at infinity (denoted by “inf”) at the right end of the graphs. However, the untruncated IPTW estimators have large variability due to a small number of observations with extremely large weights. As the weights are truncated at progressively smaller values, the variance of IPTW estimators decreases while bias increases. The impact of truncation on bias is more marked when the truncation level decreases below 20 (Figure 3, top panel).

Each simulated sample was analyzed using eight different estimators specified in Table 1 and the results are compared in Table 2. As noted in Figure 3, the untruncated estimator (Estimator 1) yields unbiased but highly variable estimates for both direct and indirect treatment effects, resulting in large RMSEs. Furthermore, the SD-to-SE ratios are much greater than 1, indicating that the SEs are seriously underestimated in the untruncated Cox MSM. Accordingly, the 95% coverage rates are much lower than the nominal 95% (first row of Table 2). Among the weight truncation approaches (Estimators 2–7), all the other approaches show a significant improvement in the SDs, SD/SE ratios, RMSEs, and 95% coverage rates over the untruncated estimators except for the approach using the untruncated estimates as the proxy (Estimator 4). This comes at the expense of a small degree of bias that increases with decreasing truncation threshold.

The methods using fixed percentile weight truncation (Estimators 2–3) and the proposed MSE-based approaches using estimates with fixed percentiles as the proxy (Estimators 5–6) yielded similar results and, on average, performed the best. Estimator 2 is the best in terms of the 95% coverage rates and Estimator 6 yields the smallest RMSE. In addition, all four percentile-based estimators (Estimators 2, 3, 5, and 6) produced more accurate estimates of SE, with the SD/SE ratios between 1.02 and 1.14. Although the results of the two-fold cross-validation approach (Estimator 7 with r = 1) are slightly worse than the methods based on percentiles; this approach is more objective, since it does not require the user to arbitrarily specify a fixed percentile or a fixed proxy for weight truncation, the optimal values of which may vary across different data structures. The performance of the CV method was improved significantly when we repeated this procedure five times, the SD/SE ratios, RMSE, and 95% coverage rates all approaching the results of percentile-based

Table 2 Comparison of IPTW estimators of the Cox MSM parameters with the weight truncation level selected using different approaches, based on 2,000 samples, with 2,000 subjects and ten observation times in each sample.

	***A*(j) (true: −0.693)**					***A*(*j* − 1) (true: −0.112)**				
Estimator[1]	**Bias × 10^2 (95% CI)**	**SD**	**SD/SE (%)**	**RMSE**	**95% cover**	**Bias × 10^2 (95% CI)**	**SD**	**SD/SE (%)**	**RMSE**	**95% cover**
1	1.4 (−0.0, 2.9)	0.334	1.5	0.335	87.2	0.0 (−1.8, 1.7)	0.398	1.56	0.398	84.6
2	4.1 (3.4, 4.8)	0.155	1.03	0.160	93.2	4.2 (3.4, 5.0)	0.174	1.05	0.179	92.2
3	4.6 (4.0, 5.2)	0.138	1.02	0.145	93.1	4.7 (4.0, 5.3)	0.154	1.05	0.161	92.0
4	2.6 (1.3, 3.9)	0.300	1.75	0.301	85.9	1.2 (−0.5, 2.8)	0.367	1.90	0.368	82.1
5	5.0 (4.4, 5.5)	0.131	1.09	0.140	91.3	4.9 (4.3, 5.6)	0.148	1.14	0.156	90.1
6	5.5 (4.9, 6.0)	0.120	1.04	0.132	91.8	5.2 (4.6, 5.8)	0.136	1.10	0.146	90.5
7, $r = 1$[2]	4.3 (3.6, 5.0)	0.156	1.26	0.162	90.7	3.9 (3.1, 4.8)	0.189	1.39	0.194	89.1
7, $r = 5$[2]	5.0 (4.5, 5.6)	0.130	1.10	0.139	91.7	4.8 (4.1, 5.5)	0.154	1.21	0.161	89.6
8	5.2 (4.8, 5.7)	0.107	0.93	0.119	94.2	4.8 (4.3, 5.3)	0.119	0.95	0.128	93.1

[1] See Table 1 for the definition of each estimator.

[2] r indicates how many times that we repeated the two-fold cross-validation approach for each sample.

methods. Although the untruncated IPTW estimator is consistent for the true causal parameter, using the untruncated IPTW estimate as the proxy produced suboptimal results (Estimator 4). Because of the large variability of untruncated IPTW estimates, the estimates from a single dataset may deviate considerably from the true parameter and thus the estimated MSE cannot be well estimated by this approach.

To further evaluate the performance of the proposed MSE-based methods (Estimators 4–8), in Figure 4 we compare the frequency distribution of the selected optimal truncation levels using different methods. The different methods yield different distributions of the selected optimal truncation levels. Except for the method with untruncated IPTW estimates as the proxy (Estimator 4), all approaches selected relatively low truncation thresholds (3–5) for the majority of the samples. For the two percentile-based proxies (Estimators 5 and 6, or "Perc99" and "Perc995" in the figure respectively), the optimal truncation levels did not exceed 15 for any sample. In contrast, the optimal truncation levels selected using two-fold cross-validation approach varied across the range of thresholds, and exhibited a pattern similar to those selected using the method based on the true parameters (Estimator 8). Using the untruncated IPTW estimates as the proxy yields, a distribution of optimal truncation levels that differs considerably from the other methods, with the highest truncation level selected in about 30% of samples.

The performance of the different methods was further assessed by calculating the probabilities of selecting the same truncation level as selected by the "gold standard" method, i.e., the MSE-based method with the true parameters (Estimator 8). Among the 2,000 samples, the probabilities for each method of selecting the same truncation level as the method using true parameters are as follows: 57.3% for "Perc99", 44.0% for "Perc995", 67.5% for "CV", and 24% for the "untruncated", indicating that the cross-validation method approximates the "gold standard" distribution slightly better than the other methods.

The performance of the proposed methods across the different sample sizes is compared in Figure 5. As the sample size increases, the RMSEs of all the models decrease; however, increasing sample size does not improve the estimation of SEs with slightly increasing SD/SE ratios. The results of the different methods are

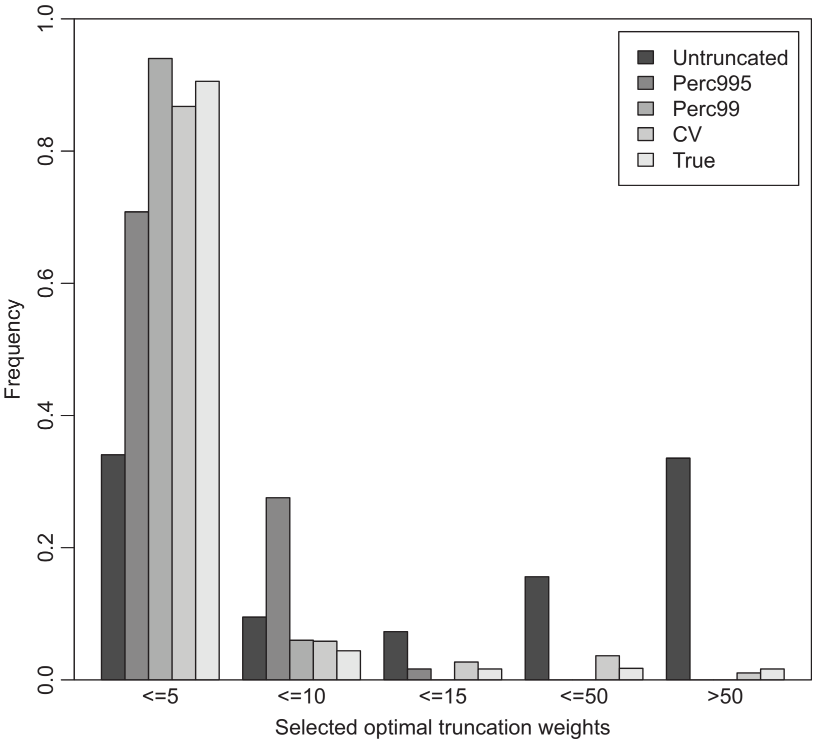

Figure 4 Distribution of the selected optimal truncation levels using different optimal truncation methods: "Untruncated", "Perc995","Perc99", "CV", and "True" corresponds to the Estimators 4–8, respectively, in Table 1.

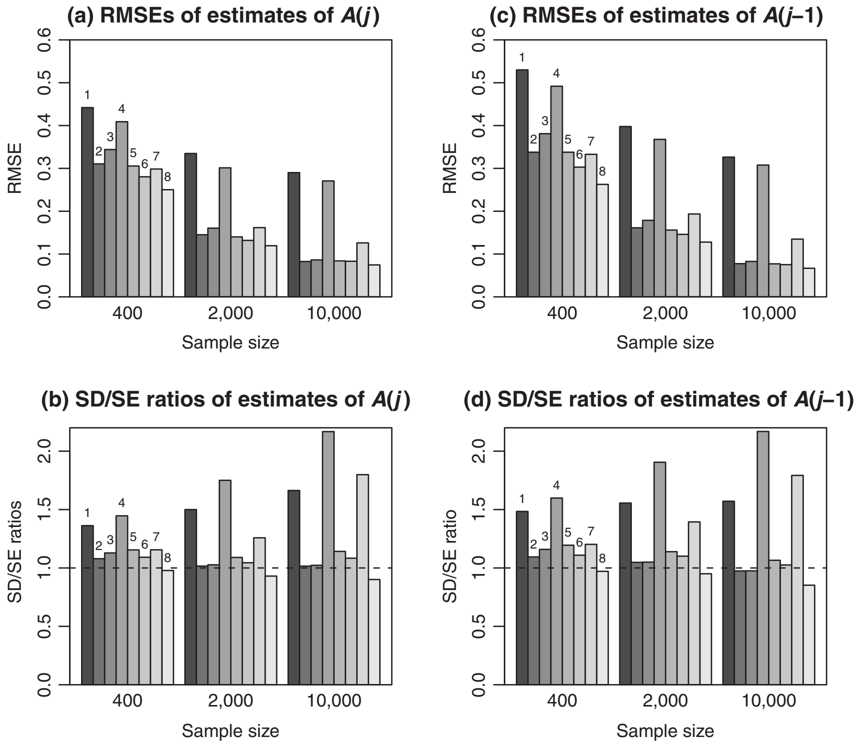

Figure 5 Comparison of the different weight truncation approaches for various sample sizes (n = 400, 2,000, 10,000). The numbers 1–8 refer to the Estimators 1–8 specified in Table 1.

similar across different sample sizes. The four percentile-based methods (Estimators 2, 3, 5, and 6) yield smaller RMSEs and more accurate variance estimates with the SD/SE ratios close to 1. The two-fold cross-validation method (Estimator 7) yields similar results to the percentile-based methods for small and medium sample size, but its performance deteriorates at the largest sample size, especially in terms of under-estimation of SE.

Finally, we investigated the performance of different methods under different degrees of violation of the ETA assumption. We found that in the scenario with only minor violation, the estimator without truncation yielded a similar RMSE to our proposed optimal methods, and the coverage rates for all the estimators was close to the nominal 95% (data not shown). In the scenario with the most serious violation, the proposed optimal methods (except for the method using the untruncated IPTW estimates as the proxies) exhibited the same performance as in the main scenario (Table 2). The method using fixed 99th percentile weight truncation yielded the best 95% coverage rate, with SD/SE ratio close to 1, while the MSE-based method using 99th percentile weight truncation IPTW estimate as the proxy gave the smallest RMSE.

4.2 Bootstrap-based SE versus Robust SE

We compared the robust sandwich SE estimator with the bootstrap-based SE estimator of the estimated treatment effects. Specifically, we randomly selected 1,000 out of 2,000 samples, and, for each sample, we generated $B = 100$ resamples by first randomly resampling subjects with replacement and then including all the observations of those selected subjects. The resamples were then analyzed using the untruncated and truncated IPTW estimators. The bootstrap-based SE was estimated as the SD of the 100 resample-specific IPTW estimates. On average, the bootstrap SEs were only slightly higher than the robust SEs, with the ratio of mean SE values equal to 1.08 for untruncated IPTW estimates (data not shown). Furthermore, for truncated estimates, the two types of SE estimators yielded very similar results (e.g., with weights truncated at 100, the ratio of the mean SE values is 1.00). However, in a small fraction of simulated samples, especially for those where the estimated treatment effect was largely affected by a single observation with an extreme weight, the bootstrap SE was occasionally much larger than the robust SE. For example, for the two samples a and b in the top panels of Figure 1, where the resample-specific estimates had a strictly bimodal distribution, the bootstrap-based SE was several times higher than the robust SE (0.524 versus 0.115, and 0.806 versus 0.211, respectively). This indicates that, in some samples, the bootstrap SE estimator is able to capture the numerical instability of the estimates that arises due to a single strongly influential outlier, whereas the robust SE estimator is not.

5 Application to the MACS

The MACS is a prospective study of the natural history of HIV infection among homosexual and bisexual men in the United States that has been ongoing since 1984. Subjects were assessed every 6 months. The information from a physical examination, laboratory results, HIV status, as well as the clinical outcomes were recorded at each visit (Kaslow et al., 1987).

Our analyses were limited to 2,002 HIV-positive men who did not have an AIDS-related disease and did not initiate AZT at the first visit between 1986 and 1992. The mean number of follow-up visits was 8 (range: 1–13). There were 511 AIDs-related deaths during the follow-up.

In our analysis, we aimed to estimate the causal effect of AZT on the time to AIDS-related death using a marginal structural Cox model (Xiao et al., 2010). Five important continuous time-dependent confounders were considered in the analysis: CD4 cell count, CD8 cell count, white blood cell count, red blood cell count, and platelet count. As only the year of a death is available in the publicly available MACS cohort data, we assigned a date of death by taking a random draw from a uniform (1,12) distribution in order to conduct a time-to-event data analysis in continuous time.

The stabilized inverse probability of treatment weights was calculated using two pooled logistic regression models that estimated the probability of being actually treated at each visit as a function of different covariates. The "denominator model" included the baseline and the current values of the time-dependent confounders, as well as the previous treatment status, whereas the "numerator model" included only the baseline values of the confounders and the previous treatment. Finally, the causal effect of AZT was estimated using a marginal structural Cox model that included terms for the use of AZT in the current and most recent 6-month period, as well as baseline values of the confounders. The weight truncation approaches described in Section 2.3 were then applied to improve the estimation in the face of large variability of the weights. For the two-fold cross-validation approach (Estimator 7), the results were averaged across 100 replications. Details of (i) specification of the treatment model used to estimate IPT weights, (ii) distribution of estimated weights, and (iii) specification of the Cox MSM used to estimate the marginal treatment effect are provided in Appendix A.2.

The estimated stabilized IPTW weights had a very large range, from 0.001 to 17,644, which resulted in the large variance of the estimates of the causal effect parameters in the marginal structural Cox model. The estimated SEs for the effects of current and previous AZT use from the marginal structural model were 0.395 and 0.476, respectively, four times higher than the corresponding SEs from the conventional multivariate Cox model (0.104 and 0.104, respectively). As shown in Table 3, except for the method using the untruncated IPTW estimates as the proxies, all the weight truncation methods reduced the SEs of the estimates significantly, with a 39–58% reduction for the estimate of the current treatment effect and 42–61% reduction for the previous treatment effect, compared with the variance of untruncated IPTW estimates. The cost for variance reduction is reflected in the change of the point estimates. The more IPT weights were truncated, the more point estimates deviated from the untruncated IPTW estimates. It is of interest to note that an excessive weights truncation may affect the statistical conclusion as well. Different from other methods, the methods based on 99th percentile (Estimators 3 and 6) affected the point estimates so much that the predictive effect of previous treatment lost statistical significance (Table 3). Our analysis showed that the current use of AZT has a nonsignificant protective effect on survival, while the most recent use of AZT has a significant protective effect on survival (except for the methods based on 99th percentile).

Table 3 Comparison of IPTW estimates of the causal effect of AZT treatment on AIDS-related mortality in the MACS, with the weight truncation level selected using different approaches.

		AZT (j)			AZT (j – 1)		
Estimator[1]	M[2]	$\hat{\beta}$	SE($\hat{\beta}$)	95% CI	$\hat{\beta}$	SE($\hat{\beta}$)	95% CI
1	–	−0.04	0.395	(−0.81, 0.74)	−1.57	0.476	(−2.50, −0.64)
2	16.2	−0.08	0.221	(−0.52, 0.35)	−0.57	0.258	(−1.08, −0.06)
3	6.9	−0.23	0.174	(−0.57, 0.12)	−0.37	0.204	(−0.77, 0.03)
4	Inf[3]	−0.04	0.395	(−0.81, 0.74)	−1.57	0.476	(−2.50, −0.64)
5	11	−0.16	0.194	(−0.54, 0.22)	−0.48	0.230	(−0.93, −0.03)
6	5	−0.24	0.164	(−0.56, 0.09)	−0.32	0.188	(−0.69, 0.05)
7[4]	22	−0.02	0.241	(−0.49, 0.45)	−0.67	0.276	(−1.22, −0.13)

[1] See Table 1 for the definition of each estimator.
[2] The selected "optimal" truncation level.
[3] Inf: indicates that the selected truncation level is not to truncate.
[4] We repeated the two-fold cross-validation approach 100 times.

As shown in the simulations study, the performance of the MSE-based methods depends on the proxy that each method used. The method using the untruncated IPTW estimates as the proxies was in favor of a larger truncation level, while the method using percentile-based proxies tended to select the optimal truncation level close to the given percentile. In contrast, the two-fold cross-validation method does not require selection of a proxy and thus yielded the most objective (data-led) results.

6 Discussion

In this paper, we conducted simulations to investigate the performance of the marginal structural Cox model when the ETA assumption is practically violated. We first illustrated the impact of ETA violation on IPTW estimation with a detailed investigation of four purposely selected poorly performing samples, which showed that poor estimation can be due to a small number of very influential observations, sometimes even a single observation, especially if it is assigned an extremely high weight. We then assessed how weight truncation can affect both the accuracy and the precision of IPTW estimators, with a particular focus on the selection of an "optimal" truncation level based on an approximated MSE. This approximate MSE calculation was accomplished by either using a proxy for the true parameter value or via a cross-validation approach.

In our main simulations, we considered a continuous confounder, a moderate confounder-treatment association, and up to ten follow-up visits such that there was a high probability of practical violation of the ETA assumption. As anticipated, although the untruncated IPTW estimators were unbiased for effects of both direct and indirect treatments, they exhibited large variability resulting in high mean-squared errors. Furthermore, the SEs of the untruncated IPTW estimates were also seriously underestimated with the SD/SE ratio greater than 1.5, resulting in the coverage rates of the 95% CIs much lower than the nominal 95%.

Cole and Hernán proposed to truncate the stabilized weights at 1% and 99% of their distribution. Our simulations show that truncating the lower tail of the weights distribution has no impact on the IPTW estimates (data not shown), and thus we considered only truncation of the upper tails. As stabilized weights were progressively truncated at lower values, the SDs for both direct and indirect treatment effect estimates decreased while the bias increased.

For the most part, the four proposed MSE-based truncated IPTW estimators as well as the two fixed-percentile truncated IPTW estimators significantly reduced the MSE of both treatment effect estimators but incurred some small bias. Weight truncation at the 99th percentile gave the most accurate estimation of SE with the SD/SE ratio almost equal to 1 and the highest coverage rates. The proposed estimator that selected the optimal truncation level by using the 99th percentile truncated weighted estimator as the proxy to approximate the expected MSE had the best performance as reflected by the lowest MSE. The performance of the two-fold CV approach was slightly worse than the methods based on percentiles; however, this approach has the advantage of being more objective in that it does not require arbitrary specification of a truncation level or a proxy. Repeating the cross-validation procedure several times significantly improved results, with MSEs approaching those of the percentile-based methods. It is interesting to note that although the overall performance of the CV approach is slightly worse than percentile-based proxy approaches, it identified the same "optimal" truncation level as the gold standard method based on the true data-generating parameters with higher frequency than all other approaches. In future research, we need to systematically investigate the impact of choice of the number of folds and the number of repetitions on the performance of the proposed cross-validation approach.

The consistent performance of the different approaches under different sample sizes and different levels of the ETA assumption violation suggests that our results and conclusions are robust with respect to sample size and the level of ETA violation.

The application of the methods to the MACS confirms that the bias-variance trade-off of the weight truncation methods and indicates that weight truncation can be used to reduce the large variance of IPTW estimates. However, we note that the excessive weight truncation may distort the estimation results. Unlike the methods that truncate the weights at a fixed level or the MSE-based methods using an user-specified proxy, the proposed two-fold cross-validation method does not require any artificial specification, and thus could be a more objective and more data-adaptive method.

Our simulations show that the large variability of the untruncated treatment effect estimator was mostly due to a small subset of samples with very unusual estimates. These estimates were, in turn, typically due to a few highly influential observations that had extreme weights resulting from unusual treatment patterns in the interval in which an event occurred. Thus, it is important to assess the influence of individual observations before reporting the IPTW estimate of a marginal structural model. Otherwise, such outliers may result in IPTW estimates that are far away from the true effect, and in extreme cases, can even reverse the direction of the effect.

In addition, our results indicate that although the average of the bootstrap-based SEs is typically very close to that of the robust SE, the bootstrap was better able to reflect instability in untruncated IPTW estimators in "bad" samples in which observations having the most extreme weights also had an event in the interval that the unusual treatment patterns were observed. Thus, the bootstrap is to be recommended to estimate the variance of IPTW estimators, although in many applications it may offer only modest improvement in the estimation of variability for truncated IPTW estimators over robust SE estimators. In addition, we found in our simulations that the SEs of IPTW estimates were much lower than their empirical SDs. Future research should address this issue.

As in 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. 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 (Sylvestre and Abrahamowicz, 2009; Vacek, 1997).

Another assumption of our main simulations was 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 the effect of treatment at the previous visit that was entirely mediated through the change in the time-dependent covariate. This critical assumption facilitated the generation of survival times conditional on the current values of the time-varying covariate and treatment, and the assessment of the accuracy of the estimates. 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 more complex data structures (Burton et al., 2006).

In addition, in the simulation studies, we assumed that both the treatment model and the marginal structural Cox model were correctly specified in the data analysis. However, in practice, any unsaturated models may be misspecified, and thus may lead to biased causal effect estimates (Robins and Hernán, 2008). Neugebauer and van der Laan (2007) developed nonparametric MSM (NPMSM) approach, which does not require correct specification of a parametric model but instead relies on a working model that can be willingly misspecified. The NPMSM may be appealing in those applications where there is no sufficient information to correctly specify the parametric model (Neugebauer and van der Laan, 2007).

Another limitation of all the methods considered in our simulations, and of other weight truncation or stabilization methods proposed in the literature, is that they will not remove bias in the case of a theoretical violation of the positivity assumption. In that case, no unusual patterns of treatment will be observed and, thus, no extreme weights will occur, so that the truncated estimates will have similar bias to untruncated ones. This limitation may also occur when the positivity assumption is *practically* violated, which is more likely to occur in small samples. When, all study subjects with a particular covariate vector receive the same treatment, no extreme weights will be assigned to this covariate pattern and, thus, weight truncation will not reduce bias in this situation.

Petersen et al. (2012) systematically reviewed alternative approaches to deal with the positivity violation and pointed out that most of these approaches represent some trade-off between unbiasedness and proximity to the initial target of inference. Alternatives to truncation include the removal of covariates that induce the most extreme weights, defining realistic treatment rules based on observed patterns in the data, or redefining the population of interest. Many of these alternative approaches rely on changing the target parameter to one that is more easily identified, which may be the only feasible solution if positivity violations are severe or by design (i.e., theoretical). The parametric bootstrap, which was proposed and validated for a point-treatment study (Wang et al., 2006), has been advocated as a tool to assess the severity of positivity violation (Petersen et al., 2012). van der Laan and Gruber (2010) developed the collaborative targeted maximum likelihood estimator (C-TMLE), in which the treatment mechanism model was data-adaptively selected in order to optimize MSE for the target parameter. The C-TMLE estimator was extended to time-to-event data by Stitelman and van der Laan (2010) and its performance was compared with alternative approaches to estimating causal effects under practical violations of the positivity assumption (Stitelman and van der Laan, 2010).

It would be interesting to investigate, in future research, whether the approaches considered in our study would improve the performance (bias and/or variance) of the IPTW estimators in the situation where estimates using untruncated weights are themselves biased (Freedman and Berk, 2008), and in situations where models are incorrectly specified.

In conclusion, our results confirm that when ETA assumption is violated, IPTW estimators of marginal structural Cox models may suffer from large variability. Simple weight truncation at high percentiles such as the 99th or the 99.5th of the distribution of weights can be applied to improve the IPTW estimators under ETA violation in most scenarios we considered. Our newly proposed data-adaptive approaches to selecting the truncation level that minimizes the expected MSE, using either observed statistics as a proxy of the true parameter or using a CV-like method, also exhibited good performance. This MSE-based method for selecting a truncation level in general can be applied to any type of estimator that relies on truncation. However, the performance of the alternative approaches should be further evaluated in a wide range of settings, including model misspecification and theoretical violations of the ETA.

A Appendix

A.1 Model specification for data generation and data analysis in simulation studies

A.1.1 Data generation procedure

For each subject $i = 1, \dots, N$, we generated the data using the following procedure:

Step 1: Generated the baseline covariate $L_i(1)$ from lognormal distribution: $L_i(1) \sim \mathit{Lognormal}\ (6, 1)$

Then for each visit $j = 1, \dots, m$, do Steps 2–4:

Step 2: Generated the treatment $A_i(j)$ from a binomial distribution with

$$\begin{aligned}&\text{logit}(P[A_i(j)|A_i(j-1), L_i(j)])\\&= 3.623 - 2605 * I[L_i(j){>}500] - 0.022 * (L_i(j) - 200)\\&+ 0.009 * (L_i(j) - 200) * I[L_i(j){>}500] + 0.405 * A_i(j-1),\end{aligned}$$

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 0 otherwise.

Step 3: Generated the confounder $L_i(j + 1)$ from a multivariable linear regression model,

$$L_i(j+1) = L_i(j) + 70A_i(j) + \Delta_i + \varepsilon_i(j+1),$$

where $\varepsilon_i(j+1) \sim N(0; \sigma = 3)$ represents sampling errors, and Δ_i was deemed to represent the yearly decline in the *ith* 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 $\Delta_i \sim U[-80, -5]$.

Step 4: Generated the expected survival time for *jth* interval, $t^\star$, using the standard inversion method from the exponential distribution with the individual interval-specific hazard rate calculated assuming the following proportional hazard model:

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

We assumed that two adjacent visits are 6 months apart. Thus, if $t^* < 0.5$ years, it indicates that the subject had an event in the interval $k = j$ and with the survival time in the last interval as $t_i^*(k)$. Otherwise, it indicates that the subject remained event-free until the end of the respective 6-month interval.

Step 5: The expected survival time was calculated as $T_i = 0.5(k-1) + t_i^*(k)$ for $1 \le k < m$, and $T_i > 5$ for $k \ge m$. The observed follow-up time for the *ith* subject was then defined as follows: $t_i = \min(T_i, C_i, 5)$ years, C_i is the censoring time, it was generated from a uniform $U[0,40]$ years to obtain about a 60% censoring rate.

A.1.2 Data analysis models

The treatment at visit j, $A_i(j)$, was estimated using a pooled logistic model conditional on the two independent variables: 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 treatment $A_i(j)$ was modeled as a function of $A_i(j-1)$ only.

The following marginal structural Cox model with adjustment for the baseline covariates was used to estimate the causal effect of $A(j)$ and $A(j-1)$:

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

The number of subjects, the percent of receiving treatment ($A = 1$), and the distribution of the estimated IPT weights of a randomly selected sample are shown in Table A.1.

Table A.1 Descriptive analysis of a generated sample across visits: the number of subjects, the treatment distribution, and the distribution of the estimated IPT weights.

Visit	*N*	*P[A = 1]*	The (untruncated) stabilized IPT weights					
			Mean	Median	Min	Max	99th	99.5th
1	2,000	0.48	0.99	0.76	0.71	28.88	5.84	8.23
2	1,814	0.52	0.94	0.66	0.43	25.46	5.29	8.40
3	1,655	0.56	0.89	0.60	0.28	28.64	5.78	7.81
4	1,527	0.55	0.86	0.53	0.15	50.71	6.25	9.83
5	1,398	0.57	0.81	0.44	0.09	39.01	6.98	9.86
6	1,285	0.57	0.82	0.37	0.04	93.45	8.72	14.33
7	1,176	0.59	6.02	0.32	0.02	6246.52	9.10	14.11
8	1,064	0.59	4.45	0.28	0.01	4081.22	7.30	9.96
9	959	0.59	4.29	0.23	0.00	3246.57	7.91	22.51
10	877	0.55	3.48	0.19	0.00	2584.40	7.25	15.20

A.2 Model specification for MACS data analysis

A.2.1 Treatment models

The treatment model, which was used to estimate the probability of receiving the AZT at a given visit m for the denominator of IPT weights, was specified using the following logistic regression model:

$$\begin{aligned}\text{logit}(P[\text{AZT}=1|m])\sim &\text{AZT}(m-1)+\text{CD4}(0)+\text{CD8}(0)+\text{RBC}(0)\\ &+\text{WBC}(0)+\text{Platelet}(0)+\text{CD4}(m)+\text{CD8}(m)\\ &+\text{RBC}(m)+\text{WBC}(m)+\text{Platelet}(m),\end{aligned}$$

where we denote Zidovudine therapy by AZT, CD4 cell count by CD4, CD8 cell count by CD8, white blood cell count by WBC, red blood cell count by RBC, and platelet count by Platelet. The model for estimating the treatment probability in the numerator of IPT weights was specified similarly as that for the denominator, except for excluding all the time-dependent variables from the model.

The distribution of the estimated stabilized IPT weights is shown in Figure A.1, in which the visit-specific IPT weights were plotted against clinical visits (every 6 months). The boxplot for each group shows the location of the median, quartiles, minimum, and maximum of logarithm of IPT weights. As expected, as the follow-up time (i.e., visit) increases, the stabilized IPT weights became more dispersed.

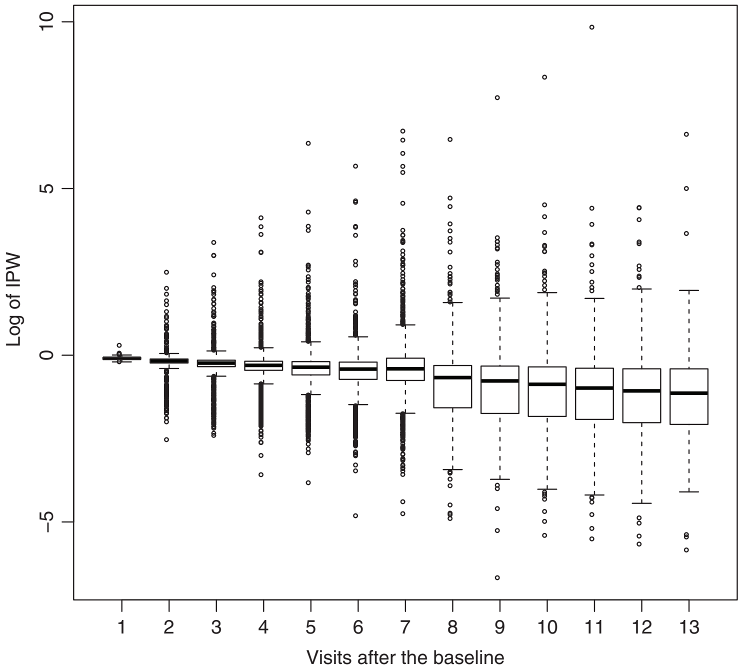

Figure A.1 Distribution of the subject-visit-specific inverse-probability-of-treatment weights over the clinical visits.

A.2.2 Marginal structural Cox model

The causal effect of AZT was estimated using a marginal structural Cox model that included two binary indicators for AZT use: the current use and the use within the most recent 6-month period, as well as baseline values of the confounders, which were used to estimate the numerator of IPT weights. Specifically, the Cox MSM was defined as a time-dependent Cox PH model, weighted with the estimated subject-visit-specific IPT weights:

$$\log(\lambda(m))\sim\text{AZT}(m)+\text{AZT}(m-1)+\text{CD4}(0)+\text{CD8}(0)+\text{RBC}(0)+\text{WBC}(0),$$

where λ (m) is the hazard at time m, AZT (m) denotes the AZT use at current visit, and AZT $(m-1)$ denotes the AZT use within the most recent 6-month period.

Acknowledgement: We wish to thank the editor and referees for their helpful suggestions. Yongling Xiao acknowledges the support of her 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 (#81275). Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS) with centers (Principal Investigators) at the Johns Hopkins Bloomberg School of Public Health (Joseph B. Margolick, Lisa P. Jacobson), Howard Brown Health Center, Feinberg School of Medicine, Northwestern University, and Cook County Bureau of Health Services (John P. Phair, Steven M. Wolinsky), University of California, Los Angeles (Roger Detels), and University of Pittsburgh (Charles R. Rinaldo). The MACS is funded by the National Institute of Allergy and Infectious Diseases, with additional supplemental funding from the National Cancer Institute. UO1-AI-35042, UL1-RR025005 (GCRC), UO1-AI-35043, UO1-AI-35039, UO1-AI-35040, UO1-AI-35041. Website located at http://www.statepi.jhsph.edu/macs/macs.html.

References

Bembom, O. and van der Laan, M. (2008a). Data-adaptive selection of the adjustment set in variable importance estimation. UC Berkeley Division of Biostatistics Working Paper Series. Paper 231.

Bembom, O. and M. van der Laan (2008b). Data-adaptive selection of the truncation level for Inverse-Probability-of-Treatment-Weighted estimators. UC Berkeley Division of Biostatistics Working Paper Series. Paper 230.

Burton, A., D. Altman, P. Royston, and R. Holder (2006). The design of simulation studies in medical statistics. Statistics in Medicine, 25(24):4279–4292.

Cole, S. and Hernán, M. (2008). Constructing inverse probability weights for marginal structural models. American Journal of Epidemiology, 168(6):656–664.

Cox, D. R. (1972). Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B, 34:187–220.

Crump, R., Hotz, V. Imbens, G. and Mitnik, O. (2006). Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. National Bureau of Economic Research, Technical Report 330.

Dehejia, R. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American statistical Association, 94(448):1053–1062.

Freedman, D. and Berk, R. (2008). Weighting regressions by propensity scores. Evaluation Review, 32(4):392–409.

Heckman, J., Ichimura, H. and Todd, P. (1998). Matching as an econometric evaluation estimator. The Review of Economic Studies, 65(2):261–294.

Hernán, M., Brumback, B. and Robins, J. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology, 11(5):561–570.

Hernán, M., Brumback, B. and Robins, J. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. Journal of the American Statistical Association, 96(454):440–448.

Kaslow, R., Ostrow, D. Detels, R. Phair, J. Polk, B. and Rinaldo Jr, C. (1987). The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. American Journal of Epidemiology, 126(2):310–318.

Kish, L. (1992). Weighting for unequal pi. Journal of Official Statistics, 8(2):183–200.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, 76(4):604– 620.

Moore, K., Neugebauer, R. and van der Laan, M. (2009). Inference in epidemiological studies with strong confounding. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper Series. Paper 255.

Neugebauer, R. and van der Laan, M. (2005). Why prefer double robust estimators in causal inference? Journal of Statistical Planning and Inference, 129(1–2):405–426.

Neugebauer, R. and van der Laan, M. (2007). Nonparametric causal effects based on marginal structural models. Journal of Statistical Planning and Inference, 137(2):419–434.

Neugebauer, R., van der Laan, M. Joffe, M. Tager, I. et al. (2007). Causal inference in longitudinal studies with history-restricted marginal structural models. Electronic Journal of Statistics, 1:119–154.

Petersen, M., Porter, K. Gruber, S. Wang, Y. and van der Laan, M. (2012). Diagnosing and responding violations in the positivity assumption. Statistical Methods in Medical Research, 21(1):31–54.

R Development Core Team (2011). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Robins, J. (1997). Marginal structural models. In:1997 Proceedings of the section on Bayesian statistical science. American Statistical Association, 1–10.

Robins, J. (1999a). Association, causation, and marginal structural models. Synthese, 121(1):151–179.

Robins, J. (1999b). Marginal structural models versus structural nested models as tools for causal inference. Statistical Models in Epidemiology: the Environment and Clinical Trials, 116:95–134.

Robins, J. and Hernán, M. (2008). Estimation of the causal effects of time-varying exposures. Longitudinal Data Analysis: A handbook of modern statistical methods. Fitzmaurice G, Davidian M, Verbeke G, et al., Chapman and Hall/CRC, New York, 553–599.

Robins, J., Hernán, M. and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology, 11(5):550–560.

Stitelman, O. and van der Laan, M. (2010). Collaborative targeted maximum likelihood for time to event data. The International Journal of Biostatistics, 6(1):21.

Sylvestre, M. and Abrahamowicz, M. (2008). Comparison of algorithms to generate event times conditional on time-dependent covariates. Statistics in Medicine, 27(14):2618–2634.

Sylvestre, M. and Abrahamowicz, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. Statistics in Medicine, 28:3437–3453.

Vacek, P. (1997). Assessing the effect of intensity when exposure varies over time. Statistics in Medicine, 16:505–513.

van der Laan, M. and Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. The international journal of biostatistics, 6(1):1–69.

van der Wal, W., Prins, M. Lumbreras, B. and Geskus, R. (2009). A simple G-computation algorithm to quantify the causal effect of a secondary illness on the progression of a chronic disease. Statistics in Medicine, 28:2325–2337.

Wang, Y., Petersen, M. Bangsberg, D. and van der Laan, M. (2006). Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. UC Berkeley Division of Biostatistics Working Paper Series. Paper 211.

Xiao, Y., Abrahamowicz, M. and Moodie, E. (2010). Accuracy of conventional and marginal structural Cox model estimators: a simulation study. The International Journal of Biostatistics, 6(2). Article 13.

Young, J., Hernán, M. Picciotto1, S. and Robins, J. (2009). Relation between three classes of structural models for the effect of a time-varying exposure on survival. Lifetime Data Analysis, 16(1):71–84.