Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Estimating inverse probability weights using super learner when weight-model specification is unknown in a marginal structural Cox model context

Mohammad Ehsanul Karim[a]*, Robert W. Platt[b,c,d,e], and The BeAMS study group[§]

Correct specification of the inverse probability weighting (IPW) model is necessary for consistent inference from a marginal structural Cox model (MSCM). In practical applications, researchers are typically unaware of the true specification of the weight model. Nonetheless, IPWs are commonly estimated using parametric models, such as the main-effects logistic regression model. In practice, assumptions underlying such models may not hold and data-adaptive statistical learning methods may provide an alternative. Many candidate statistical learning approaches are available in the literature. However, the optimal approach for a given dataset is impossible to predict. Super Learner (SL) has been proposed as a tool for selecting an optimal learner from a set of candidates using cross-validation. In this study, we evaluate the usefulness of a SL in estimating IPW in four different MSCM simulation scenarios, in which we varied the specification of the true weight model specification (linear and/or additive). Our simulations show that, in the presence of weight model misspecification, with a rich and diverse set of candidate algorithms, SL can generally offer a better alternative to the commonly used statistical learning approaches in terms of MSE as well as the coverage probabilities of the estimated effect in an MSCM. The findings from the simulation studies guided the application of the MSCM in a multiple sclerosis cohort from British Columbia, Canada (1995-2008) to estimate the impact of beta-interferon treatment in delaying disability progression. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: time-dependent confounding; marginal structural models; inverse-probability weighting; multiple sclerosis; super learner.

[a] *Centre for Health Evaluation and Outcome Sciences (CHÉOS), St. Pauls Hospital, Vancouver, BC, Canada* [b] *Department of Epidemiology, Biostatistics and Occupational Health, McGill University;* [c] *Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital;* [d] *Department of Pediatrics, McGill University;* [e] *Research Institute of the McGill University Health Centre;* [§] *'The BeAMS Study, Long-term Benefits and Adverse Effects of Beta-interferon for Multiple Sclerosis': Shirani, A.; Zhao Y.; Evans C.; Kingwell E.; van der Kop M.L.; Oger J.; Gustafson, P; Petkau, J; Tremlett, H.*

* *Correspondence to: Dr. Mohammad Ehsanul Karim, Scientist and Biostatistician, Centre for Health Evaluation and Outcome Sciences (CHÉOS), St. Pauls Hospital, 588-1081 Burrard St. Vancouver, BC, Canada, V6Z1Y6, E-mail: ekarim@cheos.ubc.ca*

Introduction

It is not always feasible to conduct a clinical trial to assess effectiveness of a treatment. Marginal structural models (MSMs) may be used to infer from an observational study that emulates a hypothetical randomized clinical trial. For survival outcomes, marginal structural Cox models (MSCMs) can be used to model the causal effect of treatment on survival in the presence of time-dependent confounding. These approaches can be illustrated using counterfactual theory [1–3]. Given potential confounding between an outcome and the treatment exposure of interest (e.g., time-dependent confounders), that could potentially distort the causal relationship of interest, methods such as inverse probability of treatment weighting (IPW), can be used obtain consistent estimates of causal effects defined by MSCMs [4–7].

IPWs are frequently used to estimate causal effects from MSCMs. The properties of the estimated IPWs influence the estimated effects from MSCM and their accuracy. As with any model, causal interpretation of a treatment effect estimate from an MSCM requires assumptions including positivity, consistency, conditional exchangeability, and correct specification of the MSCM and the weight model [7, 8]. If the weight models are correctly specified, estimates from the MSCM will be efficient. However, previous research has shown that MSCM estimates are highly sensitive to weight model misspecification [9, 10].

Few guidelines on how to calculate IPWs in a longitudinal setting are available in the literature [5, 8, 11]. They are commonly estimated using main-effects logistic regression models. In practical applications, researchers are often unaware of the true form of the weight model, i.e., whether non-linearity (e.g., quadratic or higher-order effects) or non-additivity (e.g., interaction terms) is required to describe the relationship adequately. To capture more features of the data, researchers may choose to make the parametric specification complex, i.e., include additional higher order effects and interactions. But such models may include more parameters than the observed data can support when dealing with high-dimensional data with large number of covariates or datasets with a relatively small number of observations. Arbitrary model specification can lead to erroneous inference. Assessment of the fit of logistic models for IPW is rarely seen in the MSCM literature [12, 13].

Alternative modelling strategies, such as statistical learning methods, are appealing to many as these approaches can data-adaptively detect non-linear, non-additive and higher-order effects as well as find better classification boundaries in transformed covariate spaces [14, 15]. However, it is impossible to a priori predict which approach performs best in a given dataset [16].

One proposed solution is to use Super Learner (SL) [16]. SL uses a set (library) of user-specified candidate learners or methods. This library may include parametric and semi-parametric regression models as well as data-adaptive statistical learning methods [17]. Using cross-validation, this approach optimally combines the predicted values from each candidate learner through a weighted average and computes estimated predicted values that asymptotically outperform each of the candidate estimators in the library if the correct parametric model is not included in the candidate library. Therefore, as true parametric specification of the weight model is almost always unknown, SL may offer a better alternative to logistic regression model or other data-adaptive statistical learning approaches.

As the performance of learners varies in different data sets, it is tempting to choose a rich collection of learners in the user-specified candidate library for SL. However, only a few studies have applied SL in the context of longitudinal MSM or MSCM. So far, a very limited number of learners, e.g., logistic regression models, polychotomous regressions, neural networks, k-nearest neighbors and boosted classification and regression trees (CART) have been considered as candidate learners [18–21]. Neugebauer et al. [19] studied the implementation of a SL in a single data analysis. Other studies were focused on implementing the SL algorithm in large or high-dimensional datasets and choose a limited list of candidate learners to avoid further computational burden [18, 20, 21]. There exist other studies using MSCM that were focused on data-adaptive methods, but not SL [15, 22, 23]. We extend the literature by including a wide range of diverse prediction algorithms (10 learners; described later) in the SL candidate library in a simulation study in which including a rich SL candidate library is computationally manageable. In this simulation study, for the first time, we assess the usefulness of using SL in a MSCM context in four explicit scenarios where the true parametric exposure-confounder relationships include non-additivity, non-linearity, both or none. Our aim is to assess the benefits of using super learning approach to estimate the MSCM weights in terms of estimated treatment effect MSE when the treatment assignment model is misspecified to varying degrees.

The remainder of the paper is organized as follows. In the next section, we describe the notation for MSCM, SL, design of the simulation study, and the metrics used to evaluate their performances. Then we summarize the results from the simulation scenarios and illustrate the application of SL in fitting MSCM using the data from multiple sclerosis cohort from British Columbia, Canada (1995-2008) [24, 25]. The paper concludes with a discussion of the results, and the implications and limitations of the current study.

Methods

In order to compare the performance of SL under weight model misspecification, we conduct a number of Monte Carlo simulations. Here we define the notation for MSCM and SL.

Notations for Marginal Structural Cox Model

Assume that regular measurements at visits $m = 0, 1, 2, \ldots K$ are collected in a hypothetical longitudinal study. Let the time of the baseline visit be $t_0 = 0$ and the corresponding measured covariates denoted by L_0. Let T be the exact failure time until which follow-up continues. The treatment status ($A_m = 1$ if the subject is treated in the m-th interval and $A_m = 0$ otherwise) is measured immediately after recording the value of a continuous covariate (L_m) at the m-th time interval $[t_m, t_{m+1})$. We denote $\bar{A}_m = (A_0, A_1, \ldots, A_m)$ and $\bar{L}_m = (L_0, L_1, \ldots, L_m)$ the observed treatment history and covariate history respectively through the end of interval m. The corresponding realizations of $\bar{A}_m$ and $\bar{L}_m$ are $\bar{a}_m = (a_0, a_1, \ldots, a_m)$ and $\bar{l}_m = (l_0, l_1, \ldots, l_m)$ respectively. We also define the failure indicator at the time t_{m+1} as $Y_{m+1} = I(T \leq t_{m+1})$, and the failure history through the end of interval $m + 1$ as $\bar{Y}_{m+1} = (Y_1, Y_2, \ldots, Y_{m+1})$. Let $\bar{A}_{-1} = \bar{L}_{-1} = 0$ and $Y_0 = 0$. If the time-dependent confounder $\bar{L}_m$ is a strong predictor of the treatment exposure for a subject in a given time-period, then IPW down-weights the corresponding person-time contribution. Such weighting removes the time-dependent confounding from the relationship between outcome and treatment exposure.

There are 2^{K+1} possible treatment regimes (realizations of $\bar{A}_K$): $\bar{a}_K = (a_0, a_1, \ldots a_m, \ldots a_K)$. These include $\bar{0}_K = (0, \ldots, 0)$ (never treated), $\bar{1}_K = (1, \ldots, 1)$ (always treated) and $\bar{a}_K = (0, \ldots, 0, 1, \ldots, 1)$ (partly treated) etc. We let the counterfactual failure time had a subject followed a (hypothetical) regime $\bar{a}_K$ be denoted by $T^{\bar{a}_K}$. The counterfactual outcome history under the treatment regime is, then, denoted by $\bar{Y}_{K+1}^{\bar{a}_K}$. Therefore, we can define an MSCM for regime $\bar{a}_m$ as follows:

$$\lambda_{\bar{a}_m}(m) = \lambda_{\bar{0}}(m) \exp\left(\gamma(m, \bar{a}_m, \psi_1)\right), \tag{1}$$

where the causal effect is indicated by a constant parameter vector ψ_1, γ is a known function, $\lambda_{\bar{a}_m}(m)$ and $\lambda_{\bar{0}}(m)$ are hazard functions for $T^{\bar{a}_m}$ and $T^{\bar{0}_m}$ (counterfactuals at time t_m) respectively at time m. For the treatment regime $\bar{a}_m$, the causal hazard ratio can be defined as $\lambda_{\bar{a}_m}(m)/\lambda_{\bar{0}_m}(m)$ comparing with $\bar{0}_m$. A causal effect is said to be present ($\psi_1 \neq 0$) if for any $\bar{a}_m (m = 0, 1, \ldots, K)$, $\lambda_{\bar{a}_m}(m) \neq \lambda_{\bar{0}}(m)$. The equality of the hazard functions for all K intervals is indicative of the absence of a causal effect ($\psi_1 = 0$). We can specify $\gamma = \psi_1 f(A_m) + \psi_2 L_0$ based on a function of treatment exposure (e.g., current exposure [5] or cumulative exposure [26]).

Note that, to keep notations manageable and concordant with the current literature, the above definition of the MSCM deals with the simple situation where there is only one continuous time-dependent confounder (L_m) that may affect future treatment decisions (A_m) and the hazard function under consideration ($\lambda_{\bar{a}_m}(m)$) deals with only the current treatment exposure (A_m). In practice, MSCM is capable of addressing situations where there are multiple binary, categorical or continuous time-dependent confounders and the hazard function can be fairly complicated. For example, later in this paper, we do consider a situation when two time-dependent (one binary S_m and one continuous L_m) confounders are present that affect future treatment decisions. We also consider a situation when the hazard function depends on the cumulative treatment exposure.

Estimation of ψ_1 from the MSCM

Standard modelling approaches that include L_m as a covariate, may provide biased estimates of ψ_1 if L_m is influenced by past exposure [5]. Instead of using L_m as a covariate, MSCM uses it to calculate inverse probability weights that are person-time specific measures of the degree to which L_m confounds the treatment selection process. Stabilization of the weights is generally advocated to decrease weight variation, which consequently increase the precision of MSCM estimates. The stabilized weights are derived from the following equation:

$$sw_{im} \;=\; \prod_{j=0}^{m} \frac{pr(A_{ij} = a_{ij} | \bar{A}_{i,j-1} = \bar{a}_{i,j-1}, L_{i0} = l_{i0})}{pr(A_{ij} = a_{ij} | \bar{A}_{i,j-1} = \bar{a}_{i,j-1}, L_{i0} = l_{i0}, \bar{L}_{ij} = \bar{l}_{ij})}. \tag{2}$$

Corresponding normalized version of the weight can be calculated as follows:

$$sw_{im}^{(n)} = \frac{sw_{im} n_m}{\sum_{i \in r_m} sw_{im}}, \tag{3}$$

where r_m denotes the risk-set at time m, n_m denotes the total number of subjects in the risk-set and sw and $sw^{(n)}$ are the stabilized and stabilized normalized weights respectively. These weights create a pseudo-population where the confounding due to the time-dependent confounder is removed from the relationship between outcome and treatment exposure. We fit the MSCM using the Cox model with IPWs to estimate ψ_1 in equation (1) [27] and calculate the robust sandwich standard error (calculated based on residuals and weights) [28, 29].

Estimating Inverse Probability Weights via *Super Learner*

SL starts with a candidate list of learners or prediction algorithms and evaluates the performance of each. Using V-fold cross-validation, it predicts the outcome for the validation data, based on the fit from the training data using each candidate learner. It relies on a loss function to calculate the weights for the chosen candidate learners. The optimal weighted combination of the candidate learners under consideration is the SL function that ultimately provides the SL prediction [14, 17, 21]. Web-Appendix A illustrates the steps required to implement this algorithm [14, 21]. Table 1 lists 10 candidate learners that are included in our SL's [16, 17, 30] candidate library to estimate IPWs. This list contains a wide range of learners, such as, parametric, nonparametric, tree-based and non-linear models.

All analyses were performed using `R 3.2.2` [31]. Web-Table B2 in Web-Appendix B includes the sample `R` code to fit the candidate learner models.

Table 1. Candidate learners included in the super learner under consideration.

Learner	Description	Reference
Logistic regression	The main terms of the covariates	[32]
Stepwise logistic regression	Variables selected from the main and the 2nd order terms based on AIC criterion	[33]
Elastic net	Mixing parameter = 0.5	[34]
Bayesian logistic regression	Cauchy prior with scale = 2.5	[35]
Classification and regression trees (CART)	Complexity parameter = 0.01	[36]
Pruned CART	Complexity parameter chosen such that the cross-validated error rate is minimum	[36]
Bagged CART	Based on 100 replications	[37]
Boosted CART	Based on $5,000$ trees and interaction depth = 3	[38]
Random Forest	Based on $1,000$ trees	[39]
Support vector machines	Polynomial kernel	[40, 41]

Simulating Data Suitable for MSCM Fit

The simulations performed in the current paper are extensions of the simplistic MSCM simulations proposed so far in the literature [22, 23, 27, 42–46] that were constructed based on rigorous methodological/theretical foundation [47, 48]. For example, much simulations of MSCM in the literature deal with only one binary time-dependent confounder that affects future treatment [23, 27, 42, 43, 45, 47, 48], whereas the current paper deals with a continuous time-dependent confounder, which has only recently been explored in the literature [15, 49]. In these papers, it was repeatedly shown (e.g., in [27, 47]) that, even in the simplest situation under consideration (i.e., only one binary time-dependent confounder), the MSCM fitted in the data generated from the proposed algorithm [47, 48] provides more accurate estimation than using a simple time-varying Cox model. In this study, the first simulation considers L_m to be a continuous variable so that we can assess the effect of polynomial terms in L_m and/or the interaction term $A_{m-1} \times L_m$ in the true treatment decision model (as in [15]). In particular, past treatment exposure status A_{m-1} is a predictor of L_m, which then predicts future treatment exposure A_m as well as future failure status Y_{m+1} via $1/\log(T^{\bar{0}})$. Therefore, L_m is a time-dependent confounder affecting the future treatment choices [5]. Furthermore, in the current work, we have added two more simulations to deal with scenarios with additional complexities: (i) two time dependent confounder, (continuous L_m and binary S_m) (ii) complex hazard function (cumulative treatment). A further simulation was added to assess the effect of having a larger cohort size. Web-Appendix C includes pseudocodes for MSCM data simulation. Other than the above mentioned works that follows Young et al's framework [47, 48], there exits a number of studies that have proposed various other algorithms of simulating data suitable for fitting MSCMs [11, 20, 21, 50, 51].

Simulation scenarios In this simulation scenario, L_m is a continuous variable and we assume linearity in the logit. The sampling distributions of L_m depends on its previous lagged values as well as the lagged values of L_m and A_m, i.e., l_{m-1} and a_{m-1}, as follows:

$$\begin{aligned} L_m &= E(L_m = l_m | A_{m-1}, L_{m-1}, Y_m = 0; \boldsymbol{\beta}) \\ &= \beta_0 + \beta_1 \big(1/\log(T^{\bar{0}})\big) + \beta_2 A_{m-1} + \beta_3 L_{m-1}. \end{aligned} \tag{4}$$

The time-dependent treatment A_m is sampled from a Bernoulli distribution with probability p_A. In this simulation study, to generate treatment status A_m, we considered 4 different models. The form of the true treatment models is as follows:

A. When the Hazard Function Depends on the Current Treatment Exposure :

I. Additivity and linearity: In the treatment status generating model, only main effects are included as follows:

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1 | L_m, A_{m-1}, L_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_3 L_{m-1}. \end{aligned} \quad (5)$$

II. Non-additivity: The interaction term $A_{m-1} \times L_m$ mimics the commonly occurring situation that both of these factors A_{m-1} and L_m influence future treatment decisions, which is realistic for many disease settings:

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1 | L_m, A_{m-1}, L_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_3 L_{m-1} + \alpha_4 (A_{m-1} \times L_m). \end{aligned} \quad (6)$$

III. Non-linearity: In the treatment status generating model, 2 quadratic terms are included as follows:

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1 | L_m, A_{m-1}, L_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_3 (L_{m-1})^2. \end{aligned} \quad (7)$$

IV. Non-linearity and non-additivity: The treatment status generating model includes an interaction as well as 2 quadratic terms:

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1 | L_m, A_{m-1}, L_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_3 (L_{m-1})^2 + \alpha_4 (A_{m-1} \times L_m). \end{aligned} \quad (8)$$

B. When the Hazard Function Depends on the Cumulative Treatment Exposure: In many practical scenarios, considering cumulative treatment exposure could be more realistic than considering the current treatment exposure. The history of binary treatment exposure for each patient can be cumulatively added to create the cumulative treatment exposure variable. To generate the data accordingly, the parameter for the cumulative exposure is set as the target parameter in the data generating algorithm. Unlike the previous scenarios, L_{m-1} (and consequently α_3) is not present in the following treatment generating models in order to ensure the unbiasedness of the cumulative treatment effect [15, 46].

V. Additivity and linearity: The treatment status at each stage A_m depends on the previous therapy, A_{m-1} and current disease activity, L_m:

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1 | L_m, A_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m. \end{aligned} \quad (9)$$

VI. Non-additivity: The treatment status A_m depends on the previous therapy, A_{m-1}, current disease activity, L_m and the interaction term $A_{m-1} \times L_m$:

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1 | L_m, A_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_4 (A_{m-1} \times L_m). \end{aligned} \quad (10)$$

VII. Non-linearity: The treatment status A_m depends on the previous therapy, A_{m-1} and quadratic term of the current disease activity, L_m:

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1|L_m, A_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2. \end{aligned} \quad (11)$$

VIII. Non-linearity and non-additivity: The treatment status A_m depends on the previous therapy, A_{m-1}, quadratic term of the current disease activity, L_m and the interaction term $A_{m-1} \times L_m$:

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1|L_m, A_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_4 (A_{m-1} \times L_m). \end{aligned} \quad (12)$$

C. When Multiple Time-dependent Confounders are Present that Affect Future Treatment Decisions : Previously, we generated a continuous time-dependent confounder L_m. In this new simulation scenario, we consider generating another binary time-dependent confounder, S_m. At each time interval, values of the binary S_m are sampled from a Bernoulli distribution with probability p_S, where p_S is defined as follows:

$$\begin{aligned} logit(p_S) &= logit\ Pr(S_m = 1|A_{m-1}, S_{m-1}, Y_m = 0; \boldsymbol{\beta}) \\ &= \beta_0 + \beta_1 I(T^{\bar{0}} < c) + \beta_2 A_{m-1} + \beta_3 S_{m-1}, \end{aligned} \quad (13)$$

where, $T^{\bar{0}}$ is the untreated counterfactual survival time and c is an arbitrary cut-point used to generate the binary variable $I(T^{\bar{0}} < c)$. Below we list all the treatment generating models under consideration:

IX. Additivity and linearity: Only main effects are included as follows:

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1|L_m, S_m, A_{m-1}, L_{m-1}, S_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_3 L_{m-1} + \alpha_{22} S_m + \alpha_{32} S_{m-1} \end{aligned} \quad (14)$$

X. Non-additivity: The interaction terms $A_{m-1} \times L_m$ and $A_{m-1} \times S_m$ are included:

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1|L_m, S_m, A_{m-1}, L_{m-1}, S_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_3 L_{m-1} + \alpha_4 (A_{m-1} \times L_m) \\ &\quad + \alpha_{22} S_m + \alpha_{32} S_{m-1} + \alpha_{42} (A_{m-1} \times S_m) \end{aligned} \quad (15)$$

XI. Non-linearity: 4 quadratic terms are included as follows:

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1|L_m, S_m, A_{m-1}, L_{m-1}, S_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_3 (L_{m-1})^2 + \alpha_{22} (S_m)^2 + \alpha_{32} (S_{m-1})^2 \end{aligned} \quad (16)$$

XII. Non-linearity and non-additivity: The treatment status generating model includes an interaction as well as 4 quadratic terms:

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1|L_m, S_m, A_{m-1}, L_{m-1}, S_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_3 (L_{m-1})^2 + \alpha_4 (A_{m-1} \times L_m) \\ &\quad + \alpha_{22} (S_m)^2 + \alpha_{32} (S_{m-1})^2 + \alpha_{42} (A_{m-1} \times S_m) \end{aligned} \tag{17}$$

Simulation Specifications

The true causal effect of treatment is assumed to be hazardous (ψ_1 = 0.5 on the log-hazard scale) to the subjects (in equation (1)). The associated parameter vector in equation (4) is $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (\log(3/7), -2, -\log(1/2), -\log(3/2))$. When the hazard function depends on the current treatment exposure, the associated parameter vectors in equations (5)-(8) are $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4) = (\log(2/7), (1/2), (-1/2), -\log(3/5), 1.2)$. When the hazard function depends on the cumulative treatment exposure, the associated parameter vectors in equations (9)-(12) are $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_4) = (\log(2/7), (1/2), (-1/2), 1.2)$. When multiple time-dependent confounders are present that affect future treatment decisions, the associated parameter vectors in equations (14)-(17) are $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_{22}, \alpha_{32}, \alpha_{42}) = (\log(2/7), (1/2), (-1/2), -\log(3/5), 1.2, 0.5, 1.1, -0.8)$. Also, The associated parameter vector in equation (13) is $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_{32}) = (\log(3/7), -2, -\log(1/2), -\log(3/2), 2.5)$.

We have generated cohorts with $n = 250$ subjects, each with up to 10 visits, to assess the impact of estimating the MSCM parameter ψ_1 using weights via SL. The event rate under consideration is $\lambda_0 = 0.10$ (monthly events throughout the follow-up). To assess characteristics of a larger cohort, these simulations are repeated for $n = 1,500$ subjects, with maximum 10 follow-up visits. The Monte Carlo study consists of $N = 1,000$ simulated datasets for each setting under consideration.

Specification of Weight Models to Estimate IPW

The numerator model for the stabilized weights (shown in equation (2)) included the lagged value of treatment status A_{m-1} and the follow-up month index m. The denominator model for the stabilized weights included the numerator model covariates as well as the time-dependent covariate L_m for simulation settings V-VIII. For simulation settings I-IV, the lagged value L_{m-1} was also included in the denominator model. For simulation settings IX-XII, we further included the time-dependent covariate S_m and the lagged value S_{m-1} in the denominator model. In a given simulation setting, while estimating IPW via any candidate learners as well as SL, we used the same covariate list. For example, in simulation setting XII, while estimating denominator model of IPW via any candidate learners, we included the following covariates: A_{m-1}, L_m, L_{m-1}, S_m, S_{m-1} and m. In contrast, in simulation setting VIII, the following covariates were used: A_{m-1}, L_m and m to estimate denominator model of IPW via any candidate learners.

Performance metrics

We assessed the performance of the weighting schemes using the following measures

- Bias $= \sum_{q=1}^{N} (\hat{\psi}_1 - \psi_1)/N$: The average difference between the true and $N = 1000$ estimated parameter (log hazards ratio) from the MSCM model. Here, $q = 1, 2, \ldots, N = 1000$.
- SD $= \sqrt{\sum_{q=1}^{N} (\hat{\psi}_1 - \bar{\psi}_1)^2/(N-1)}$ where $\bar{\psi} = \sum_{q=1}^{N} \hat{\psi}_1/N$
- MSE $= \sqrt{\sum_{q=1}^{N} (\hat{\psi}_1 - \psi_1)^2/N}$

- Model-based SE: The average of $N = 1000$ estimated standard errors of the estimated causal effect from the MSCM model.
- Coverage probabilities of model-based nominal 95% CIs: Proportion of $N = 1000$ datasets in which true parameter was contained in the estimated 95% CI.

Results

A. Simulation Settings I - IV:

a) Stabilized Weights and Cohort Size, $n = 250$: The results obtained from the simulation scenarios I-IV (when the hazard function depends on the current treatment exposure) are presented in Tables 2 - 5. The methods are listed in ascending order with respect to the estimated MSE. In general the SL approach did well in all of these simulation settings as it includes both parametric and non-parametric approaches as candidate learners. Among the candidate learners, boosted CART performed very well in all scenarios. The MSE of this method was closest to that of the SL approach. One exception was when the treatment model was linear and additive. Then elastic net, which is known to be more stable than logistic regression, did slightly better than the boosted CART approach. In general, SVM and random forest performed poorly in our simulations in terms of MSE and coverage probabilities. Below we outline the performance of parametric and non-parametric candidate learners in terms od bias, SD and MSE.

When the treatment generating model was additive and linear (simulation - I: main-effects only) and candidate learners also used an additive and linear form for the weight model specification, parametric models, such as elastic net, Bayesian logistic regression and logistic regression performed well in terms of bias. These parametric models generally ranking lower than the CART methods (w.r.t. bias) when we considered non-additive terms (simulation - II that includes interaction) or non-linear terms (simulation - III that includes polynomials) or both (simulation - IV that includes interaction and polinomial terms) in the treatment generating model. In general CART methods can automatically consider interaction and polynomial terms and it is not surprising that these methods perform better than the parametric models in the presence of non-additive or non-linear terms.

In terms of SD, parametric models only performed well in the simulation - I scenario. In simulation - II scenario, bagged and boosted CART methods perform better (w.r.t. SD) than these parametric models. As soon as a non-linear term is introduced (simulation - III and IV), parametric models perform worse (w.r.t. SD) that even the inferior CART models (simple and pruned CARTS).

While considering MSE, we can see the parametric models perform similarly to superior CART methods (boosted and bagged CARTs) only for simulation setting - I (see Table 2). For rest of the settings, these parametric models generally do perform worse (see Tables 3, 4, 5).

b) Normalized Stabilized weights: As a sensitivity analysis, we used $sw^{(n)}$ instead of sw in the MSCMs. The corresponding results are presented in Web-Table D1-D4 in Web-Appendix D. The results are, in principle, similar to those when sw was used in the MSCMs.

c) Truncated Stabilized Weights: We also applied 1% truncation to the stabilized weights (sw). The corresponding results are presented in Web-Table E1-E4 in Web-Appendix E. As shown in these tables, when the large weights were truncated as little as 1%, MSE of resulting effect estimate reduced. This resulted in logistic regression and elastic net to do better compared to the untruncated analysis in terms of MSE in the scenarios where non-linear terms were present

Table 2. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is linear and additive in describing the association between the exposure and the confounder (simulation scenario - I, with weights sw)

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	-0.0719	0.0844	0.312	0.281	0.969
Elastic net	-0.1336	0.1031	0.308	0.292	0.934
Boosted CART	-0.1493	0.1039	0.314	0.286	0.951
Bayesian logistic	0.0195	0.1071	0.323	0.327	0.972
Logistic	0.0645	0.1218	0.329	0.343	0.972
Bagged CART	-0.2469	0.2749	0.386	0.463	0.837
Stepwise	0.1458	0.3750	0.346	0.595	0.950
CART	-0.4232	0.4221	0.397	0.493	0.722
Pruned CART	-0.6215	0.6246	0.342	0.488	0.507
SVM	0.3807	1.7024	0.502	1.248	0.601
Random Forest	-0.6002	2.4178	0.309	1.434	0.148

Table 3. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is linear but non-additive in describing the association between the exposure and the confounder (simulation scenario - II, with weights sw)

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	0.00825	0.0312	0.185	0.176	0.970
Boosted CART	0.02492	0.0316	0.187	0.176	0.965
Bagged CART	-0.00614	0.0325	0.193	0.180	0.965
Stepwise	0.03801	0.0654	0.223	0.253	0.966
Random Forest	0.03017	0.0741	0.294	0.270	0.973
CART	0.02451	0.0769	0.229	0.276	0.914
Pruned CART	-0.04692	0.0849	0.222	0.288	0.867
Elastic net	0.21918	0.0881	0.207	0.200	0.839
Bayesian logistic	0.24436	0.1011	0.210	0.203	0.822
Logistic	0.25562	0.1083	0.213	0.207	0.814
SVM	0.16572	0.2104	0.240	0.428	0.845

Table 4. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is additive but non-linear in describing the association between the exposure and the confounder (simulation scenario - III, with weights sw)

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	0.1059	0.259	0.468	0.498	0.9667
CART	0.2217	0.324	0.473	0.524	0.9170
Bagged CART	0.3249	0.342	0.491	0.486	0.9410
Boosted CART	0.3369	0.357	0.498	0.493	0.9157
Pruned CART	-0.0903	0.411	0.472	0.634	0.8550
Stepwise	0.2332	0.594	0.421	0.735	0.7655
Elastic net	0.3813	0.601	0.540	0.675	0.8960
Bayesian logistic	0.4147	0.602	0.571	0.656	0.9000
Logistic	0.3290	0.695	0.488	0.766	0.8847
Random Forest	-1.1129	1.402	0.434	0.405	0.2420
SVM	2.1906	6.128	0.323	1.153	0.0287

in the true weight model. In terms of MSE, pruned CART did not do well. As before, random forest and SVM still did not perform well. Web-Appendix I includes the graphs (Figures I1-I12) of these summaries (e.g., bias, MSE, coverage

Table 5. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is non-linear and non-additive in describing the association between the exposure and the confounder (simulation scenario - IV, with weights sw)

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	0.0952	0.241	0.468	0.481	0.9688
CART	0.2038	0.305	0.471	0.513	0.9230
Bagged CART	0.3018	0.315	0.487	0.473	0.9470
Boosted CART	0.3386	0.356	0.497	0.491	0.9137
Pruned CART	-0.1086	0.402	0.470	0.624	0.8560
Bayesian logistic	0.3522	0.506	0.562	0.618	0.9190
Elastic net	0.3490	0.551	0.537	0.655	0.9060
Stepwise	0.2186	0.555	0.428	0.712	0.7928
Logistic	0.3110	0.660	0.490	0.751	0.8933
Random Forest	-1.1216	1.423	0.434	0.406	0.2320
SVM	2.1899	6.137	0.324	1.158	0.0289

probability) when weights are progressively truncated at higher percentiles in each simulation scenario. In general, the bias of all weight estimation approaches agreed at 50% truncation. This matches with the expectation that when weights are truncated 50%, irrespective of the weight estimation technique, the corresponding treatment effect estimate converges to that from a baseline-adjusted (i.e., unadjusted for time-dependent confounding) analysis [8].

d) Stabilized Weights and Larger Cohort Size, $n = 1,500$: The results from the simulations when larger cohorts are available (i.e., n = 1,500) are shown in Web-Tables F1-F4. The ψ_1 estimates perform much better in terms of SD in all the settings under consideration compared to the n = 250 case (Tables 2-5). Consequently, the MSEs are generally smaller in all settings with larger cohort sizes. When cohort size increases, the parametric models generally show better performance in terms of bias than the CART methods (simple, bagged and pruned) compared to the scenarios with smaller cohort size, except for the scenario where non-additive terms are included in the treatment generating model. The boosted CART method perform very well in terms of bias in all scenarios. In terms of SD, the bagged and boosted CART models show better performance (w.r.t. SD) than the parametric methods, except for the scenario where non-linear terms are included in the treatment generating model. In general, in terms of MSE, SL and boosted CART methods perform better than most of the other candidate learners in the settings we have considered. In particular, boosted CART method performed best in simulation settings - I, II, and IV, whereas SL performed best in simulation setting - III. However, in this larger cohort scenario, when we compare these two methods (SL and boosted CART) in terms of 95% coverage probability, it is apparent that coverage probabilities from SL are always closer to 0.95 compared to that from the boosted CART. However, irrespective of the cohort sizes, a few characteristics remain the same: e.g., stepwise (which includes second order interaction terms) and CART methods are generally doing better in simulation setting - II (that includes interaction or non-additive term) and SVM and random forest methods generally perform worse in most settings.

B. Simulation Settings V - VIII:

When the Hazard Function Depends on the Cumulative Treatment Exposure: Web-Tables G1-G4 shows the pattern of MSE in estimating the MSCM parameter when the hazard function depends on the cumulative treatment exposure instead of current treatment exposure. Except for simulation setting VI, stepwise method performs best in terms of bias. However, boosted CART performs best in terms of SD. Only in simulation setting V, boosted CART outperforms SL in terms of MSE. However, when compared with respect to 95% coverage probability in the same setting (simulation setting V), coverage probability from SL is actually closer to 0.95 compared to that from the boosted CART. For rest of the settings (simulation settings VI, VII and VIII), SL outperforms all the candidate learners under consideration in terms of MSE. When the true treatment selection model includes non-linearity in the covariates (scenarios VII and VIII), we can see increased magnitudes of bias associated with the mis-specified IPW estimation methods under consideration. This potentially explains the poor coverage probabilities for all methods in the corresponding scenarios (scenario VII and VIII).

C. Simulation Settings IX - XII:

When Multiple Time-dependent Confounders are Present that Affect Future Treatment Decisions: Web-Tables H1-H4 shows the pattern of MSE for the MSCM parameter when two time-dependent confounders are present. In simulation settings IX and X, both boosted and bagged CART methods perform better than the SL in terms of MSE. Unlike previous cases, the coverage probabilities of the SL approach are also further away from 0.95 compared to both of these methods (boosted and bagged CART). For simulation settings XI and XII, SL outperforms these two methods, along with the other candidate learners under consideration in terms of bias, SD and MSE.

Multiple Sclerosis Data Analysis

We apply the methodology described in this paper to the British Columbia (BC) MS cohort data (1995-2008) [15, 23–25, 52–57] to estimate the effect of β-IFN on time to irreversible disability progression. In this study, irreversible progression of disability is measured by sustained expanded disability status scale (EDSS) 6, which is confirmed after at least 150 days, with all subsequent EDSS scores being 6 or greater. Web-Appendix §J describes the baseline characteristics, eligibility and exclusion criteria of the MS cohort. Based on these criteria, $1,697$ patients were included in the study [24, 25]. At the end of follow-up, 138 subjects reached irreversible disability, measured by sustained Expanded Disability Status Scale (EDSS) of score 6.

In this analysis, 'cumulative relapses over the last two years' (hereafter 'cumulative relapses') was considered as a time-dependent confounder. β-IFN exposure was defined as a time-dependent variable A_m, measured on a monthly basis. MSCMs are an appropriate choice of model to adjust for the time-dependent confounder L_m cumulative relapses and baseline confounders L_0: age, sex, disease duration, and EDSS score [25]. To estimate the stabilized weights sw, we used the SL with the same candidate learners used in our simulations. For all learners, the numerator model for treatment and censoring included the baseline covariates L_0 (EDSS score, age, disease duration, sex), the lagged treatment status A_{m-1}, and the follow-up month index m. The denominator model included the numerator model covariates as well as the time-dependent covariate L_m 'cumulative relapses'. The resulting sw weighted MSCM further adjusted for the baseline covariates to estimate the hazard ratio ($\hat{HR} = \exp(\hat{\psi_1}) = 1.35$) and corresponding confidence intervals based on robust standard errors (0.316; see Table 6). Web-Table K1 in Web-Appendix K shows corresponding estimates from the fitted MSCMs for increased levels of weight truncation. In general, the weights were well-behaved (mean approximately 1 and small SD) and the analyses did not yield any strong evidence of an association between β-IFN exposure and time to reaching sustained EDSS score of 6.

Discussion

When estimating weights for the MSCM, it is common practice to use main-effects logistic regression to model the treatment decision process. However, misspecification of the weight model (e.g., when the model deviates from linearity or additivity) severely affects the estimate of the MSCM treatment effect. Our aim is to see if the uses of super learning approach to estimate the weights can improve the MSE and the coverage probabilities of the treatment effect estimate

Table 6. The marginal structural Cox model (MSCM) fit with the normalized stabilized inverse probability of treatment and censoring weights sw for time to sustained EDSS 6 to estimate the causal effect of β-IFN treatment for multiple sclerosis (MS) patients from British Columbia, Canada (1995-2008). The model was also adjusted for the baseline covariates EDSS, age, disease duration and sex.

Covariate	Estimate*	HR †	SE(HR)	95% CI ‡
β-IFN	0.30	1.35	0.32	0.85 - 2.13
EDSS	0.48	1.62	0.14	1.36 - 1.92§
Disease duration#	-0.22	0.81	0.16	0.59 - 1.10
Age#	0.26	1.30	0.13	1.01 - 1.68§
Sex¶	-0.29	0.75	0.19	0.45 - 1.23

HR, Hazard ratio; CI, confidence interval; EDSS, expanded disability status scale

* Estimated log HR

† HR, indicating the instantaneous risk of reaching sustained and confirmed EDSS 6

‡ Based on robust standard error.

§ 95% CI that does not include 1.

# Expressed in decades.

¶ Reference level: Male

when the treatment assignment model is misspecified.

We assessed the performance of the SL approach with a diverse list of candidate learners for estimating IPW in a MSCM context via simulation. We considered four settings characterized by varying degrees of deviance from linearity and additivity to describe the treatment decision model. When stabilized weights were computed via this SL, the resulting MSCM estimates computed from SL generally performed better in terms of MSE and the coverage probabilities compared to individual candidate learners. It harnessed the power of both parametric and non-parametric approaches to produce better results in all the scenarios under consideration. Similar performances were also observed when stabilized normalized weights were computed from the same SL. These simulations shows the utility of using SL approach with rich set of candidate learners in practical scenarios when the form of the treatment decision model is unknown and may deviate from linearity, additivity or both.

Throughout the simulation scenarios, one of the candidate learners, boosted CART, performed very close to that of SL. This approach was shown to outperform logistic regression as well as other popular statistical learning approaches in its use in the propensity score context [58] and in the context of estimation of MSCM weights [15]. SVMs performed poorly in the current context, which is also not surprising [15]. It is unclear how SVMs correct for confounding variables in their predictions. Use of confounder-correcting SVM (ccSVM) may aid in improving predictions in the presence of confounders [59]. When the SL candidate library includes such poorly performing learners in a given context, it is possible that the learners with good performance, such as bagged and boosted CART, may outperform SL in some complex scenarios in terms of MSE. However, even in those scenarios, SL generally performs close to the best performing learner in terms of MSE or sometimes performs even better in terms of coverage probabilities.

Based on the overall good performance of the SL approach in our simulations, we implemented a SL with the same rich candidate library in a MS application. We estimated the effect of β-IFN on irreversible disability progression using the MSCM with a stabilized weights estimated via the SL. The hazard ratio estimates from the super learning approach is

1.349, and this effect estimate was not significant (95% CI $0.853 - 2.134$). This conclusion is consistent with those of the previous studies [24, 25].

We have also assessed the performance of the practice of weight truncation. From previous literature, we know that small amount of weight truncation is often helpful and a suggested practice in the literature when large weights are present [15, 22]. Not surprisingly, truncation at a lower level (e.g., 1%) improved the performance of parametric approach such as logistic regression model in the current context. However, as shown in the previous literature, there is a bias-variance trade-off associated with such practice and researchers need to be cautious about using this approach [8]. For example, when the weights were generated from the logistic regression approach and subsequently 5% truncated, the estimated HR was reported as 1.11 (95% CI $0.64 - 1.95$) in the same MS dataset [25]. Although the corresponding CI width was similar to the one obtained in the current study using a SL approach, the HR substantially moved towards the null as a result of truncation.

This work has some caveats. The ability of SL approach depends on the choice of candidate learners, and hence results from our simulation cannot be generalized beyond the settings and candidate learners we considered. The performance of the candidate learners also depends on tuning the parameters chosen by the analyst. In our study, we mostly retained the default settings offered by off-the-shelf package `SuperLearner` [60], thereby facilitating use of these methods by practitioners. The SL approach is an ensemble method, that requires fitting all the candidate learners considered in the user-specified library in order to obtain the final SL prediction. Therefore, the computational burden of SL is generally much higher than that of standard approaches such as logistic regression. This is especially true when computationally intensive learners, such as bagged CART or boosted CART [15] are included in the candidate library [20]. In general, the computation time for SL is at least twice the sum of all the candidate learners' computation time, considering fitting on the training sets, computing the corresponding weights from the validation sets and fitting the entire data eventually [61]. Similar to other MSCM simulation studies [22, 27], we computed robust sandwich standard error [28, 29] in this paper. To get more reliable estimate of the standard error, resampling methods, such as the bootstrap could be used [8, 42, 62].

Many of the statistical learning approaches used as candidate learners for the SL library are very useful variable-selection tools for choosing the variables that need to be adjusted for in the model as well as identifying the functional form of their empirical relationship. However, these tools are not meant to replace subject-matter knowledge and expert-opinion. For example, in a treatment modelling context, if we control for variables that are unrelated to outcome, although it can improve the predictive performance, the efficiency of the effect estimate may suffer [63–66]. Moreover, if we control for known instrumental variables and colliders, the bias may be amplified [67]. However, recent simulation studies have shown that, such increases in bias are rather small in practical epidemiological settings compared to bias resulting from omitting confounders from the analysis [68–70]. In practical situations, when it may be hard to determine whether a variable is a confounder or an instrumental variable, these studies suggested that it is possibly more harmful to omit a variable (i.e., under-adjustment) rather than controlling for it (i.e, over-adjustment). However, there may be other specific settings when bias amplification may be substantial [71]. In our simulations, we assumed 'no unmeasured confounders' and did not consider any instrumental variables. Future simulation studies emulating practical settings could examine whether these obstacles have a major impact in the context of MSCM weight computation.

Acknowledgements

Funding Information

This work was supported by a post-doctoral fellowship from the Canadian Network for Observational Drug Effect Studies (CNODES) (M.E.K.). R.W.P. holds the Albert Boehringer I Chair in Pharmacoepidemiology and is supported in part by a National Scholar (Chercheur-national) of the Fonds de Récherche du Québec-Santé (FQR-S) and is a member of the Research Institute of the McGill University Health Centre, which is supported by core funds from FQR-S.

Conflict of Interest

M.E.K. has received accommodation costs from the endMS Research and Training Network (2011, 2012), Statistical Society of Canada (2016) to present at conferences, and from Pacific Institute for the Mathematical Sciences (2013), The Canadian Statistical Sciences Institute (2016) to attend workshops. R.W.P. has received fees for service for consulting from Abbvie, Amgen, and Searchlight Pharma, for teaching from Novartis, and for scientific steering committee membership from Pfizer. A.S. has received travel grants to present at and attend conferences from the endMS Research and Training Network (2010, 2011), ECTRIMS (2010, 2011), and the Consortium of MS Centres (2012). Y.Z. declares no conflicts of interest. C.E. has received travel grants to present and attend conferences from the endMS Research and Training Network (2011), and the European Committee for Treatment and Research in MS (2011). E.K. has had travel and accommodation costs covered to present at and attend conferences from the endMS Research and Training Network (2008, 2011), the International Society for Pharmacoepidemiology (2010, 2013) and Bayer Schering Pharma (2010). M.L.vdK. declares no conflicts of interest. J.O. receives fees for services from Bayer, from Novartis and from Biogen Idec to serve on advisory committees. Over the past five years, J.O. has received speaker honoraria, consulting fees, travel grants and/or research grants and/or educational grants from Aventis, Bayer, Biogen-Idec, BioMS, Corixa, Genentech, Novartis, Serono, Shering, Talecris and Teva-neurosciences. P.G. has received consulting fees from Biogen to serve on an advisory board. Over the past three years, J.P. has received consulting fees and/or fees for service on Data Safety Monitoring Boards from EMD Serono, Myelin Research Foundation and Novartis. H.T. has received speaker honoraria and/or travel expenses to attend conferences from the Consortium of MS Centres (2013), the National MS Society (2012, 2014, 2016), Teva Pharmaceuticals (2011), ECTRIMS (2011, 2012, 2013, 2014, 2015, 2016), UK MS Trust (2011), the Chesapeake Health Education Program, US Veterans Affairs (2012), Novartis Canada (2012), Biogen Idec (2014), American Academy of Neurology (2013, 2014, 2015, 2016). All speaker honoraria are either declined or donated to an MS charity or to an unrestricted grant for use by her research group.

Thank you's

We gratefully acknowledge the British Columbia Multiple sclerosis Clinic neurologists who contributed to the study through patient examination and data collection (current members listed here by primary clinic): UBC MS Clinic: Dr. A. Traboulsee, MD, FRCPC (UBC Hospital MS Clinic Director and Head of the UBC MS Programs); Dr. A-L. Sayao, MD, FRCPC; Dr. V. Devonshire, MD, FRCPC; Dr. S. Hashimoto, MD, FRCPC (UBC and Victoria MS Clinics); Dr. J. Hooge, MD, FRCPC (UBC and Prince George MS Clinics); Dr. L. Kastrukoff, MD, FRCPC (UBC and Prince George MS Clinics); Dr. J. Oger, MD, FRCPC. Kelowna MS Clinic: Dr. D. Adams, MD, FRCPC; D. Craig, MD, FRCPC; Dr. S. Meckling, MD, FRCPC. Prince George MS Clinic: Dr. L. Daly, MD, FRCPC. Victoria MS Clinic: Dr. O. Hrebicek, MD, FRCPC; Dr. D. Parton, MD, FRCPC; Dr. K Pope, MD, FRCPC. The views expressed in this paper do not necessarily reflect the views of each individual acknowledged. None have received compensation for their role in the study.

References

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

[2] Hernán M. Invited Commentary: Hypothetical Interventions to Define Causal EffectsAfterthought or Prerequisite? *American Journal of Epidemiology* 2005; **162**(7):618, doi:10.1093/aje/kwi255.

[3] Morgan S, Winship C. *Counterfactuals and causal inference*. Cambridge University Press, 2014.

[4] Robins J. Marginal structural models versus structural nested models as tools for causal inference. *Statistical Models in Epidemiology, the Environment and Clinical Trials* 1999; **116**:95–134.

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

[6] Cole S, Hernán M, Margolick J, Cohen M, Robins J. Marginal structural models for estimating the effect of highly active antiretroviral therapy initiation on CD4 cell count. *American Journal of Epidemiology* 2005; **162**(5):471–478.

[7] Platt R, Brookhart M, Cole S, Westreich D, Schisterman E. An information criterion for marginal structural models. *Statistics in medicine* 2013; **32**(8):1383–1393.

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

[9] Mortimer K, Neugebauer R, Van der Laan M, Tager I. An application of model-fitting procedures for marginal structural models. *American Journal of Epidemiology* 2005; **162**(4):382–388.

[10] Lefebvre G, Delaney J, Platt R. Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statistics in Medicine* 2008; **27**(18):3629–3642.

[11] Bryan J, Yu Z, van der Laan M. Analysis of longitudinal marginal structural models. *Biostatistics* 2004; **5**(3):361–380.

[12] Suarez D, Borras R, Basagana X. Differences between marginal structural models and conventional models in their exposure effect estimates: A systematic review. *Epidemiology* 2011; **22**(4):586–588.

[13] Yang S, Eaton C, Lu J, Lapane K. Application of marginal structural models in pharmacoepidemiologic studies: a systematic review. *Pharmacoepidemiology and Drug Safety* 2014; **23**(6):560–571.

[14] Rose S. Mortality risk score prediction in an elderly population using machine learning. *American journal of epidemiology* 2013; **177**(5):443–452.

[15] Karim M, Petkau J, Gustafson P, Tremlett H, BeAMS. On the application of statistical learning approaches to construct inverse probability weights in marginal structural cox models: Hedging against weight-model misspecification 2016. URL `http://www.tandfonline.com/doi/abs/10.1080/03610918.2016.1248574`, communications in Statistics-Simulation and Computation. DOI: 10.1080/03610918.2016.1248574 (published online: 21 October).

[16] van der Laan M, Polley E. Super learner. *Statistical Applications in Genetics and Molecular Biology* 2007; **6**(1):1–23.

[17] Pirracchio R, Petersen M, van der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *American journal of epidemiology* 2014; :kwu253.

[18] Neugebauer R, Chandra M, Paredes A, J Graham D, McCloskey C, S Go A. A marginal structural modeling approach with super learning for a study on oral bisphosphonate therapy and atrial fibrillation. *Journal of Causal Inference* 2013; **1**(1):21–50.

[19] Neugebauer R, Fireman B, Roy JA, Raebel MA, Nichols GA, O'Connor PJ. Super learning to hedge against incorrect inference from arbitrary parametric assumptions in marginal structural modeling. *Journal of clinical epidemiology* 2013; **66**(8):S99–S109.

[20] Neugebauer R, Schmittdiel JA, Zhu Z, Rassen JA, Seeger JD, Schneeweiss S. High-dimensional propensity score algorithm in comparative effectiveness research with time-varying interventions. *Statistics in medicine* 2014; **34**(5):753–781.

[21] Gruber S, Logan RW, Jarrín I, Monge S, Hernán MA. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Statistics in medicine* 2015; **34**(1):106–117.

[22] Xiao Y, Moodie E, Abrahamowicz M. Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods* 2012; **2**(1):1–20.

[23] Karim ME. Causal Inference Approaches for Dealing with Time-Dependent Confounding in Longitudinal Studies, with Applications to Multiple Sclerosis Research. PhD Thesis, University of British Columbia, Vancouver January 2015.

[24] Shirani A, Zhao Y, Karim M, Evans C, Kingwell E, van der Kop M, Oger J, Gustafson P, Petkau J, Tremlett H. Association between use of interferon beta and progression of disability in patients with relapsing-remitting multiple sclerosis. *Journal of American Medical Association* 2012; **308**(3):247–256.

[25] Karim ME, Gustafson P, Petkau J, Zhao Y, Shirani A, Kingwell E, Evans C, van der Kop M, Oger J, Tremlett H. Marginal Structural Cox Models for Estimating the Association Between β-Interferon Exposure and Disease Progression in a Multiple Sclerosis Cohort. *American Journal of Epidemiology* 2014; **180**(2):160–171.

[26] Cole S, Jacobson L, Tien P, Kingsley L, Chmiel J, Anastos K. Using marginal structural measurement-error models to estimate the long-term effect of antiretroviral therapy on incident AIDS or death. *American Journal of Epidemiology* 2010; **171**(1):113–122.

[27] Xiao Y, Abrahamowicz M, Moodie E. Accuracy of conventional and marginal structural Cox model estimators: A simulation study. *The International Journal of Biostatistics* 2010; **6**(2):1–28.

[28] Lin D, Wei L. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* 1989; **84**(408):1074–1078.

[29] Binder D. Fitting Cox's proportional hazards models from survey data. *Biometrika* 1992; **79**(1):139–147.

[30] van der Laan M, Rose S. *Targeted learning: causal inference for observational and experimental data*. Springer, 2011.

[31] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2015. URL `https://www.R-project.org/`.

[32] McCullagh P, Nelder JA. *Generalized linear models.* London England Chapman and Hall 1983., 1989.

[33] Chambers JM, Hastie TJ. *Statistical models in S*. CRC Press, Inc., 1991.

[34] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 2010; **33**(1):1.

[35] Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2008; :1360–1383.

[36] Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and regression trees*. CRC press, 1984.

[37] Breiman L. Bagging predictors. *Machine Learning* 1996; **24**(2):123–140.

[38] Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *Journal of Animal Ecology* 2008; **77**(4):802–813.

[39] Breiman L. Random forests. *Machine learning* 2001; **45**(1):5–32.

[40] Cortes C, Vapnik V. Support-vector networks. *Machine learning* 1995; **20**(3):273–297.

[41] Chang C, Lin C. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2011; **2**(3):1–27.

[42] Ali R, Ali M, Wei Z. On computing standard errors for marginal structural Cox models. *Lifetime Data Analysis* 2014; **20**(1):106–131.

[43] Young J, Tchetgen Tchetgen E. Simulation from a known Cox MSM using standard parametric models for the g-formula. *Statistics in Medicine* 2014; **33**(6):1001–1014.

[44] Moodie E, Stephens D, Klein M. A marginal structural model for multiple-outcome survival data: assessing the impact of injection drug use on several causes of death in the canadian co-infection cohort. *Statistics in Medicine* 2014; **33**(8):1409–1425.

[45] Mojaverian N, Moodie EE, Bliu A, Klein MB. The impact of sparse follow-up on marginal structural models for time-to-event data. *American journal of epidemiology* 2015; :kwv152.

[46] Xiao Y, Abrahamowicz M, Moodie E, Weber R, Young J. Flexible marginal structural models for estimating the cumulative effect of a time-dependent treatment on the hazard: reassessing the cardiovascular risks of didanosine treatment in the swiss hiv cohort study. *Journal of the American Statistical Association* 2014; **109**(506):455–464.

[47] Young J, Hernán M, Picciotto S, Robins J. Relation between three classes of structural models for the effect of a time-varying exposure on survival. *Lifetime Data Analysis* 2010; **16**(1):71–84.

[48] Young J, Hernán M, Picciotto S, Robins J. Simulation from structural survival models under complex time-varying data structures. *JSM Proceedings, Section on Statistics in Epidemiology*, American Statistical Association, 2008; 1–6.

[49] Vourli G, Touloumi G. Performance of the marginal structural models under various scenarios of incomplete marker's values: A simulation study. *Biometrical Journal* 2015; **57**(2):254–270.

[50] Westreich D, Cole S, Schisterman E, Platt R. A simulation study of finite-sample properties of marginal structural Cox proportional hazards models. *Statistics in Medicine* 2012; **31**(19):2098–2109.

[51] Havercroft W, Didelez V. Simulating from marginal structural models with time-dependent confounding. *Statistics in Medicine* 2012; **31**(30):4190–4206.

[52] Shirani A, Zhao Y, Karim M, Evans C, Kingwell E, van der Kop M, Oger J, Gustafson P, Petkau J, Tremlett H. Investigation of heterogeneity in the association between interferon beta and disability progression in multiple sclerosis: an observational study. *European Journal of Neurology* 2014; **21**(6):835–844.

[53] Zhang T, Shirani A, Zhao Y, Karim M, Gustafson P, Petkau J, Evans C, Kingwell E, van der Kop M, Zhu F, *et al.*. Beta-interferon exposure and onset of secondary progressive multiple sclerosis. *European Journal of Neurology* 2015; **22**(6):990–1000.

[54] Shirani A, Zhao Y, Petkau J, Gustafson P, , Karim M, Evans C, Kingwell E, van der Kop M, Oger J, *et al.*. Multiple sclerosis in older adults: the clinical profile and impact of interferon beta treatment. *BioMed Research International* 2015; **2015**(451912):1–11.

[55] Karim M, Gustafson P, Petkau J, Tremlett H, BeAMS. Comparison of statistical approaches for dealing with immortal time bias in drug effectiveness studies. *American Journal of Epidemiology* 2016; **184**(4):325–335.

[56] Karim M, Petkau J, Gustafson P, Platt R, Tremlett H, BeAMS. Comparison of statistical approaches dealing with time-dependent confounding in drug effectiveness studies 2016. URL `http://smm.sagepub.com/content/early/2016/09/21/0962280216668554`, statistical Methods in Medical Research. DOI: 10.1177/0962280216668554 (published online: September 21).

[57] Karim M, Gustafson P, Petkau J, Tremlett H. The authors reply. *American Journal of Epidemiology* 2016; **184**(11):857–858, doi:10.1093/aje/kww158.

[58] Lee B, Lessler J, Stuart E. Weight trimming and propensity score weighting. *PLoS one* 03 2011; **6**(3):e18 174.

[59] Li L, Rakitsch B, Borgwardt K. ccsvm: correcting support vector machines for confounding factors in biological data classification. *Bioinformatics* 2011; **27**(13):i342–i348.

[60] Polley E, van der Laan M. *SuperLearner: Super Learner Prediction* 2014. URL `http://CRAN.R-project.org/package=SuperLearner`, r package version 2.0-15.

[61] Ju C, Combs M, Lendle SD, Franklin JM, Wyss R, Schneeweiss S, van der Laan MJ. Propensity score prediction for electronic healthcare dataset using super learner and high-dimensional propensity score method 2016. URL `http://biostats.bepress.com/ucbbiostat/paper351/`, U.C. Berkeley Division of Biostatistics Working Paper Series. Last accessed: 9th Dec, 2016.

[62] Austin P. Variance estimation when using inverse probability of treatment weighting (iptw) with survival analysis. *Statistics in Medicine* 2016; **35**(30):5642–5655.

[63] Rubin D. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* 1997; **127**(8 Part 2):757–763.

[64] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *American journal of epidemiology* 2006; **163**(12):1149–1156.

[65] Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in medicine* 2007; **26**(4):734–753.

[66] Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and drug safety* 2011; **20**(3):317–320.

[67] Wyss R, Stürmer T. Balancing automated procedures for confounding control with background knowledge. *Epidemiology (Cambridge, Mass.)* 2014; **25**(2):279.

[68] Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM, Glynn RJ. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology* 2011; :kwr364.

[69] Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Glynn RJ. Myers et al. respond to “understanding bias amplification”. *American journal of epidemiology* 2011; :kwr353.

[70] Liu W, Brookhart MA, Schneeweiss S, Mi X, Setoguchi S. Implications of m bias in epidemiologic studies: a simulation study. *American journal of epidemiology* 2012; **176**(10):938–948.

[71] Pearl J. Invited commentary: understanding bias amplification. *American journal of epidemiology* 2011; **174**(11):1223–1227.

Estimating inverse probability weights using super learner when weight-model specification is unknown in a marginal structural Cox model context

Mohammad Ehsanul Karim, Robert W. Platt, and The BeAMS study group

Supporting Information:

A. Steps for Super Learner

The aim of any statistical model is to represent the true data generating process. It is common practice to rely on parametric regression methods. In a real word data analysis, in the absence of knowledge about the true data generating process, parametric regressions may be too restrictive and are unlikely to adequately represent complex relationships. Flexible semiparametric or nonparametric models have some compelling advantages in this regard. With nonparametric approaches, such as statistical learning methods, the analyst does not have to commit to a specific functional form for the relationship between the variables in a given dataset. Tree-based statistical learning methods (e.g., classification and regression trees (CART), bagged CART and boosted CART), have the ability to discover non-linear, polynomial and interaction terms. On the other hand, support vector machines (SVMs) are better equipped to solve classification problems by mathematically transforming variables into higher-dimensional spaces. However, the performance of various statistical learning methods varies in different datasets and analyst will not know a priori which learner (or method) to choose a priori. Ensembling and the Super Learner (SL) allow the analyst to combine multiple learners into one single learner and obtain the optimal prediction measured by a chosen criterion (minimum cross-validated mean squared error is typically used).The analyst can strengthen the predictive power of the Super Learner by including various types of candidate learners (parametric regression, flexible and nonparametric statistical learners).

The steps for implementing SL are as follows:

1. First we need to choose the candidate learners that we want to use in our SL. This step has to be done priori to any data analysis. For the purpose of illustration, let us choose $k = 5$ candidate learners (e.g., choose 5 learners from Table 1), say, Logistic regression, elastic net, CART, random forest and SVM.
2. We need to decide how many cross-validation sub-samples or blocks we want to use. For the purpose of illustration, let us chose $V = 3$ and let that we have $n = 90$ participants or subjects in the data. The sub-samples should contain an approximately equal number of subjects or samples ($n/V = 90/3 = 30$ subjects in each sub-sample), be mutually exclusive (non-overlapping), randomly allocated and should be similar in distribution with respect to the two treatment classes. These sub-samples will be used for performing V-fold cross validation.
3. We use V-fold cross-validation ($V = 3$) to obtain the predicted probabilities for each learner to fill the prediction matrix ($n \times K$) shown in Table SA1. To fill each column, we perform the following for each learner ($k = 5$):
   $i.$ In order to get these predictions, we first use first 2 sub-samples (‘training sets’) to fit each learner, where treatment A is the dependent variable and covariates (say, L and the corresponding lag value) that influences the treatment generation process are considered as independent variables. This fit is evaluated in the remaining sub-sample (here, the 3rd sub-sample is the ‘validation set’). This way we calculate predicted probabilities for the validation set.

   $ii.$ Next we define the 1st and 3rd sub-samples as training sets and evaluate the fit on the 2nd sub-sample. We calculate predicted probabilities for 2nd sub-sample.

   $iii.$ Finally, we define the last 2 sub-samples as training set and evaluate the fit on the 1st sub-sample. We calculate predicted probabilities for 1st sub-sample.

For each given learners, the above three steps will provide us the predicted probabilities for all subjects in all sub-samples (i.e., step i will provide predicted probabilities from the 3rd sub-sample, step ii will provide from the 2nd sub-sample and step iii will provide from the 1st sub-sample). That means, for $n = 90$ subjects in the data with 30 subjects in each sub-sample, for the first learner (logistic regression), step i will provide predicted probabilities for 30 subjects from the 3rd sub-sample, step ii will provide predicted probabilities for 30 participants from the 2nd sub-sample and step iii will provide predicted probabilities for 30 subjects from the 1st sub-sample. In the prediction matrix ($n \times K$) shown in Table SA1, these three segments (of predicted probabilities from sub-samples 1, 2 and 3) will constitute the column 1 of predicted probabilities (for all 90 subjects) obtained from the logistic regression. Now, we move on to the second learner (elastic net), and obtain the entire column 2 of predicted probabilities (for all 90 participants) obtained from the elastic net approach. Similarly we fill the column $3 - 5$ using the rest of the learners (e.g., CART, random forest and SVM). Therefore, for $k = 5$ candidate learners, we will have 5 columns of predicted probabilities $p_k; k = 1, 2, \dots, 5$.

4. We run a binary regression $P(A = 1|P) = expit(\eta_1 p_1 + \eta_2 p_2 + \eta_3 p_3 + \eta_4 p_4 + \eta_5 p_5)$ where treatment status A is considered as the dependent variable and 5 columns of predicted probabilities p_k are considered as independent variables. We estimate η_k. To increase the stability of the SL, we add the restriction that $\sum(\eta_k) = 1$ and $\eta_k \geq 0$. Let that, we have, $\eta_1 = .3$, $\eta_2 = 0$, $\eta_3 = .6$, $\eta_4 = .1$ and $\eta_5 = 0$. We retain only the η's associated with non-zero coefficients.
5. Fit all $k = 5$ algorithms on the entire dataset. We obtain 5 columns of predicted probabilities $\bar{Q}_k; k = 1, 2, \dots, 5$.
6. To get the SL prediction, we use the columns of step 5 ($\bar{Q}_k$) weighted by the coefficients of step 4 (η_k): $\bar{Q}_{SL} = \eta_1 \bar{Q}_1 + \eta_3 \bar{Q}_3 + \eta_4 \bar{Q}_4$. Candidate learners 2 and 5 were omitted as their corresponding η's were zero. Incorporating the prediction from an algorithm with zero coefficient does not contribute to substantial improvement of the overall fit. According to the chosen loss function (minimum expected squared error), this weighted combination should be associated with the smallest cross-validated mean-squared error (CV MSE; i.e., average of sum of squares of the difference of treatment status A and the predicted probabilities).

Table S A1. Prediction matrix ($n \times K$) in step 3 of the super learner estimation. In this example, sample size, $n = 90$, number of learners under consideration, $K = 5$ and number of sub-samples, $V = 3$.

Sub-sample	Subject id	Column 1 Logistic	Column 2 Elastic net	Column 3 CART	Column 4 Random forest	Column 5 SVM
Sub-sample 1	subject 1	-[†]	-	-	-	-
	subject 2	-	-	-	-	-
	...	-	-	-	-	-
	subject 30	-	-	-	-	-
Sub-sample 2	subject 31	-	-	-	-	-
	subject 32	-	-	-	-	-
	...	-	-	-	-	-
	subject 60	-	-	-	-	-
Sub-sample 3	subject 61	-	-	-	-	-
	subject 62	-	-	-	-	-
	...	-	-	-	-	-
	subject 90	-	-	-	-	-

[†] Each of the blank cells will be filled with a predicted value calculated by the learner described at the top of that column via V-fold cross-validation.

B. Example Code

Table SB2 includes the sample R code to fit the candidate learner models in our SL:

Table S B2. R codes for the chosen candidate learners included in the super learner.

Learner	Sample code	Package
Logistic regression	glm(Y ∼ ., data = X, family = binomial("logit"))	stats
Stepwise logistic regression	fit = glm(Y ∼ ., data = X, family = binomial("logit")); step(fit, scope = Y ∼ .⋀2, direction = "both", trace = 0, k = 2)	stats
Elastic net	cv.glmnet(x = X, y = Y, nfolds = 10, alpha = 0.5, nlambda = 100, binomial("logit"))	glmnet
Bayesian logistic regression	bayesglm(Y ∼ ., data = X, family = binomial("logit")	arm
CART	rpart(Y ∼ ., data = data.frame(Y, X), control = rpart.control(cp = 0.01, minsplit = 20, xval = 10, maxdepth = 30, minbucket = round(minsplit/3)))	rpart
Pruned CART	fit=rpart(Y ∼ ., data = data.frame(Y, X), control = rpart.control(cp = 0.001, minsplit = 20, xval = 10, maxdepth = 20, minbucket = 5));mincp = fit $cptable [which.min(fit.rpart $cptable[, "xerror"]), "CP"];prune(fit, cp = mincp)	rpart
Bagged CART	ipredbagg(y=Y, X = X, nbagg = 100, control = rpart.control(xval = 0, maxsurrogate = 0, minsplit = 20, cp = 0.01, maxdepth = 30))	ipred
Boosted CART	gbm(formula, data = X, distribution = "bernoulli", n.trees = 5000, interaction.depth = 3, cv.folds = 10, keep.data = TRUE, n.minobsinnode = 10,shrinkage=0.01, bag.fraction = 1.0,train.fraction = 1)	gbm
Random Forest	randomForest(y = as.factor(Y), x = X, ntree = 1000, keep.forest = TRUE, nodesize = 1, mtry = max(floor(ncol(X)/3), 1))	randomForest
Support vector machines	svm(y = as.factor(Y), x = X, type.class = "C−classification", nu = 0.5, fitted = FALSE, probability = TRUE, kernel = "polynomial")	e1071

C. Pseudo-code for MSCM simulation

Pseudocodes used for MSCM data generation in our simulations are as follows:

C.1. Hazard Function Depends on the Current Treatment Exposure

```
**GET** $K \leftarrow 10$ (maximum follow-up);
  $\lambda_0 \leftarrow 0.10$;
  $n \leftarrow 250$ (and $1,500$ for larger cohorts);
  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) \leftarrow [\log(3/7), -2, -\log(1/2), -\log(3/2)]$ (parameter vector for generating $L$);
  $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4) \leftarrow [\log(2/7), (1/2), (-1/2), -\log(3/5), 1.2]$ (parameter vector for generating $A$);
  $\psi_1 \leftarrow 0.5$ (true log-hazard value of the treatment effect)

**COMPUTE** FOR $ID$ = 1 to $n$
  INIT: $L_{-1} \leftarrow 0$; $A_{-1} \leftarrow 0$; $Y_0 \leftarrow 0$; $H_m \leftarrow 0$
  $T^{\bar{0}} \sim$ Exponential$(\lambda_0)$
  FOR $m$ = 0 to $K$
    $L_m \leftarrow E(L_m = l_m | L_{m-1}, A_{m-1}, Y_m = 0; \beta)$
      $\leftarrow \beta_0 + \beta_1 \big(1/\log(T^{\bar{0}})\big) + \beta_2 A_{m-1} + \beta_3 L_{m-1}$
    $logit\ p_A \leftarrow$ logit $Pr(A_m = 1 | L_m, L_{m-1}, A_{m-1}, Y_m = 0; \alpha)$
      $\leftarrow \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_3 L_{m-1}$
        (for simulation scenario - I: main effects only: additive and linear)
      $\leftarrow \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_3 L_{m-1} + \alpha_4 (A_{m-1} \times L_m)$
        (for simulation scenario - II: non-additive effect: a 2-way interaction)
      $\leftarrow \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_3 (L_{m-1})^2$
        (for simulation scenario - III: non-linear effects: 2 quadratic terms)
      $\leftarrow \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_3 (L_{m-1})^2 + \alpha_4 (A_{m-1} \times L_m)$
        (for simulation scenario - IV: non-additive & non-linear effects)
    $A_m \sim$ Bernoulli$(p_A)$
    $H_m \leftarrow \int_0^{m+1} \lambda_{\bar{a}_j}(j) dj$
      $\leftarrow H_m + \exp(\psi_1 A_m)$
    IF $T^{\bar{0}} \geq H_m$
      $Y_{m+1} \leftarrow 0$
    ELSE
      $Y_{m+1} \leftarrow 1$
      $T \leftarrow m + (T^{\bar{0}} - H_m) \times \exp(-\psi_1 A_m)$
    END IF
  ENDFOR $m$
ENDFOR $ID$

**PRINT** $ID, m, Y_{m+1}, A_m, L_m, A_{m-1}, L_{m-1}$
```

C.2. Hazard Function Depends on the Cumulative Treatment Exposure

```
**GET** $K \leftarrow 10$ (maximum follow-up);
  $\lambda_0 \leftarrow 0.10$;
  $n \leftarrow 250$;
  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) \leftarrow [\log(3/7), -2, -\log(1/2), -\log(3/2)]$ (parameter vector for generating $L$);
  $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_4) \leftarrow [\log(2/7), (1/2), (-1/2), 1.2]$ (parameter vector for generating $A$);
  $\psi_1 \leftarrow 0.5$ (true log-hazard value of the cumulative treatment effect)

COMPUTE    FOR $ID$ = 1 to $n$
  INIT: $L_{-1} \leftarrow 0$; $A_{-1} \leftarrow 0$; $Y_0 \leftarrow 0$; $H_m \leftarrow 0$; $cum(A_{-1}) \leftarrow 0$.
  $T^{\bar{0}} \sim$ Exponential$(\lambda_0)$
  FOR $m$ = 0 to $K$
    $L_m \leftarrow E(L_m = l_m | L_{m-1}, A_{m-1}, Y_m = 0; \beta)$
      $\leftarrow \beta_0 + \beta_1 \big(1/\log(T^{\bar{0}})\big) + \beta_2 A_{m-1} + \beta_3 L_{m-1}$
    $logit\ p_A \leftarrow$ logit $Pr(A_m = 1 | L_m, L_{m-1}, A_{m-1}, Y_m = 0; \alpha)$
      $\leftarrow \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m$
        (for simulation scenario - V: main effects only: additive and linear)
      $\leftarrow \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_4 (A_{m-1} \times L_m)$
        (for simulation scenario - VI: non-additive effect: a 2-way interaction)
      $\leftarrow \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2$
        (for simulation scenario - VII: non-linear effects: 2 quadratic terms)
      $\leftarrow \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_4 (A_{m-1} \times L_m)$
        (for simulation scenario - VIII: non-additive & non-linear effects)
    $A_m \sim$ Bernoulli$(p_A)$
    $cum(A_m) \leftarrow cum(A_{m-1}) + A_m$
    $H_m \leftarrow H_m + \exp(\psi_1 cum(A_m))$
    IF $T^{\bar{0}} \geq H_m$
      $Y_{m+1} \leftarrow 0$
    ELSE
      $Y_{m+1} \leftarrow 1$
      $T \leftarrow m + (T^{\bar{0}} - H_m) \times \exp(-\psi_1 cum(A_m))$
    END IF
  ENDFOR $m$
ENDFOR $ID$

PRINT    $ID$, $m$, $Y_{m+1}$, $A_m$, $L_m$, $A_{m-1}$, $L_{m-1}$, $cum(A_m)$
```

C.3. Multiple Time-dependent Confounders Present that Affect Future Treatment Decisions

```
GET    $K \leftarrow 10$ (maximum follow-up);
  $\lambda_0 \leftarrow 0.10$;
  $n \leftarrow 250$;
  $c \leftarrow 30$;
  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_{32}) \leftarrow [\log(3/7), -2, -\log(1/2), -\log(3/2), 2.5]$ (parameter vector for generating $L$ and $S$);
  $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_{22}, \alpha_{32}, \alpha_{42}) \leftarrow [\log(2/7), (1/2), (-1/2), -\log(3/5), 1.2, 0.5, 1.1, -0.8]$ (parameter vector for generating $A$);
  $\psi_1 \leftarrow 0.5$ (true log-hazard value of the treatment effect)

COMPUTE    FOR $ID$ = 1 to $n$
  INIT: $L_{-1} \leftarrow 0$; $A_{-1} \leftarrow 0$; $Y_0 \leftarrow 0$; $H_m \leftarrow 0$
  $T^{\bar{0}} \sim$ Exponential$(\lambda_0)$
  FOR $m$ = 0 to $K$
    $logit\ p_S \leftarrow$ logit $Pr(S_m = 1|S_{m-1}, A_{m-1}, Y_m = 0; \beta)$
      $\leftarrow \beta_0 + \beta_1 I(T^{\bar{0}} < c) + \beta_2 A_{m-1} + \beta_{32} S_{m-1}$
    $S_m \sim$ Bernoulli$(p_S)$
    $L_m \leftarrow E(L_m = l_m|L_{m-1}, A_{m-1}, Y_m = 0; \beta)$
      $\leftarrow \beta_0 + \beta_1 \big(1/\log(T^{\bar{0}})\big) + \beta_2 A_{m-1} + \beta_3 L_{m-1}$
    $logit\ p_A \leftarrow$ logit $Pr(A_m = 1|L_m, L_{m-1}, S_m, S_{m-1}, A_{m-1}, Y_m = 0; \alpha)$
      $\leftarrow \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_3 L_{m-1} + \alpha_{22} S_m + \alpha_{32} S_{m-1}$
         (for simulation scenario - IX: main effects only: additive and linear)
      $\leftarrow \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_3 L_{m-1} + \alpha_{22} S_m + \alpha_{32} S_{m-1} + \alpha_4 (A_{m-1} \times L_m) + \alpha_{42} (A_{m-1} \times S_m)$
         (for simulation scenario - X: non-additive effect: 2-way interactions)
      $\leftarrow \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_3 (L_{m-1})^2 + \alpha_{22} (S_m)^2 + \alpha_{32} (S_{m-1})^2$
         (for simulation scenario - XI: non-linear effects: 4 quadratic terms)
      $\leftarrow \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_3 (L_{m-1})^2 + \alpha_{22} (S_m)^2 + \alpha_{32} (S_{m-1})^2 + \alpha_4 (A_{m-1} \times L_m) + \alpha_{42} (A_{m-1} \times S_m)$
         (for simulation scenario - XII: non-additive & non-linear effects)
    $A_m \sim$ Bernoulli$(p_A)$
    $H_m \leftarrow \int_0^{m+1} \lambda_{\bar{a}_j}(j) dj$
      $\leftarrow H_m + \exp(\psi_1 A_m)$
    IF $T^{\bar{0}} \geq H_m$
      $Y_{m+1} \leftarrow 0$
    ELSE
      $Y_{m+1} \leftarrow 1$
      $T \leftarrow m + (T^{\bar{0}} - H_m) \times \exp(-\psi_1 A_m)$
    END IF
  ENDFOR $m$
ENDFOR $ID$

**PRINT**   $ID, m, Y_{m+1}, A_m, L_m, A_{m-1}, L_{m-1}$
```

D. Using Normalized Stabilized Weights

Tables SD1-SD4 shows the summaries of MSCM results when $sw^{(n)}$ weights were used in the MSCMs:

Table S D1. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is linear and additive in describing the association between the exposure and the confounder (simulation scenario - I, with weights swn)

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	-0.0719	0.0844	0.312	0.281	0.969
Elastic net	-0.1336	0.1031	0.308	0.292	0.934
Boosted CART	-0.1493	0.1039	0.314	0.286	0.951
Bayesian logistic	0.0195	0.1071	0.323	0.327	0.972
Logistic	0.0645	0.1218	0.329	0.343	0.972
Bagged CART	-0.2469	0.2749	0.386	0.463	0.837
Stepwise	0.1458	0.3750	0.346	0.595	0.950
CART	-0.4232	0.4221	0.397	0.493	0.722
Pruned CART	-0.6215	0.6246	0.342	0.488	0.507
SVM	0.3807	1.7024	0.502	1.248	0.601
Random Forest	-0.6002	2.4178	0.309	1.434	0.148

Table S D2. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is linear but non-additive in describing the association between the exposure and the confounder (simulation scenario - II, with weights swn)

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	0.00825	0.0312	0.185	0.176	0.970
Boosted CART	0.02492	0.0316	0.187	0.176	0.965
Bagged CART	-0.00614	0.0325	0.193	0.180	0.965
Stepwise	0.03801	0.0654	0.223	0.253	0.966
Random Forest	0.03017	0.0741	0.294	0.270	0.973
CART	0.02451	0.0769	0.229	0.276	0.914
Pruned CART	-0.04692	0.0849	0.222	0.288	0.867
Elastic net	0.21918	0.0881	0.207	0.200	0.839
Bayesian logistic	0.24436	0.1011	0.210	0.203	0.822
Logistic	0.25562	0.1083	0.213	0.207	0.814
SVM	0.16572	0.2104	0.240	0.428	0.845

Table S D3. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is additive but non-linear in describing the association between the exposure and the confounder (simulation scenario - III, with weights swn)

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	0.1059	0.259	0.468	0.498	0.9667
CART	0.2217	0.324	0.473	0.524	0.9170
Bagged CART	0.3249	0.342	0.491	0.486	0.9410
Boosted CART	0.3369	0.357	0.498	0.493	0.9157
Pruned CART	-0.0903	0.411	0.472	0.634	0.8550
Stepwise	0.2332	0.594	0.421	0.735	0.7655
Elastic net	0.3813	0.601	0.540	0.675	0.8960
Bayesian logistic	0.4147	0.602	0.571	0.656	0.9000
Logistic	0.3290	0.695	0.488	0.766	0.8847
Random Forest	-1.1129	1.402	0.434	0.405	0.2420
SVM	2.1906	6.128	0.323	1.153	0.0287

Table S D4. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is non-linear and non-additive in describing the association between the exposure and the confounder (simulation scenario - IV, with weights swn)

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	0.0952	0.241	0.468	0.481	0.9688
CART	0.2038	0.305	0.471	0.513	0.9230
Bagged CART	0.3018	0.315	0.487	0.473	0.9470
Boosted CART	0.3386	0.356	0.497	0.491	0.9137
Pruned CART	-0.1086	0.402	0.470	0.624	0.8560
Bayesian logistic	0.3522	0.506	0.562	0.618	0.9190
Elastic net	0.3490	0.551	0.537	0.655	0.9060
Stepwise	0.2186	0.555	0.428	0.712	0.7928
Logistic	0.3110	0.660	0.490	0.751	0.8933
Random Forest	-1.1216	1.423	0.434	0.406	0.2320
SVM	2.1899	6.137	0.324	1.158	0.0289

E. Using 1% Truncated Stabilized Weights

Tables SE1-SE4 shows the summaries of MSCM results when we applied 1% truncation on the stabilized weights (sw):

Table S E1. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is linear and additive in describing the association between the exposure and the confounder (simulation scenario - I, with 1 percent truncated weights sw)

	Bias	MSE	SE	SD	Cov.Pr.
Logistic	-0.209	0.0953	0.274	0.227	0.928
Stepwise	-0.166	0.0972	0.282	0.264	0.937
Bayesian logistic	-0.243	0.1087	0.271	0.223	0.895
Super learner	-0.294	0.1377	0.273	0.226	0.855
Boosted CART	-0.327	0.1725	0.287	0.256	0.830
Elastic net	-0.357	0.1758	0.265	0.221	0.758
Bagged CART	-0.330	0.2467	0.347	0.371	0.791
CART	-0.490	0.4119	0.358	0.414	0.653
SVM	-0.634	0.5986	0.337	0.444	0.480
Pruned CART	-0.734	0.7432	0.311	0.452	0.401
Random Forest	-0.667	2.5115	0.304	1.438	0.152

Table S E2. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is linear but non-additive in describing the association between the exposure and the confounder (simulation scenario - II, with with 1 percent truncated weights swn)

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	-0.03361	0.0283	0.176	0.165	0.960
Boosted CART	-0.00255	0.0286	0.181	0.169	0.967
Bagged CART	-0.04414	0.0307	0.186	0.169	0.964
Stepwise	-0.00675	0.0347	0.199	0.186	0.968
CART	-0.02322	0.0510	0.209	0.225	0.945
SVM	0.05271	0.0549	0.229	0.228	0.935
Elastic net	0.16668	0.0613	0.193	0.183	0.876
Random Forest	0.02297	0.0696	0.290	0.263	0.975
Bayesian logistic	0.18973	0.0699	0.195	0.184	0.858
Logistic	0.19881	0.0741	0.197	0.186	0.846
Pruned CART	-0.10244	0.0797	0.207	0.263	0.851

Table S E3. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is additive but non-linear in describing the association between the exposure and the confounder (simulation scenario - III, with with 1 percent truncated weights swn)

	Bias	MSE	SE	SD	Cov.Pr.
Bagged CART	0.01709	0.171	0.482	0.413	0.990
Logistic	-0.08342	0.193	0.504	0.431	0.990
Elastic net	-0.03482	0.201	0.532	0.447	0.989
Boosted CART	0.05679	0.215	0.484	0.460	0.969
Super learner	-0.24995	0.215	0.456	0.391	0.969
Bayesian logistic	0.00167	0.222	0.557	0.471	0.987
CART	-0.04451	0.279	0.471	0.526	0.935
Stepwise	0.18140	0.380	0.558	0.589	0.928
SVM	0.08047	0.404	0.534	0.630	0.900
Pruned CART	-0.36151	0.523	0.465	0.627	0.793
Random Forest	-1.13148	1.444	0.434	0.405	0.232

Table S E4. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is non-linear and non-additive in describing the association between the exposure and the confounder (simulation scenario - IV, with with 1 percent truncated weights swn)

	Bias	MSE	SE	SD	Cov.Pr.
Bagged CART	-0.00119	0.163	0.477	0.404	0.993
Logistic	-0.10034	0.191	0.499	0.426	0.990
Elastic net	-0.05451	0.196	0.524	0.440	0.989
Bayesian logistic	-0.03656	0.208	0.541	0.454	0.988
Boosted CART	0.05727	0.218	0.482	0.463	0.965
Super learner	-0.25664	0.219	0.454	0.391	0.968
CART	-0.06114	0.272	0.469	0.518	0.938
Stepwise	0.13323	0.348	0.547	0.574	0.940
SVM	0.06720	0.442	0.530	0.662	0.900
Pruned CART	-0.37529	0.522	0.462	0.617	0.792
Random Forest	-1.14088	1.466	0.434	0.406	0.221

F. Using Larger Cohorts

Tables SF1-SF4 shows the summaries of MSCM results when we applied the stabilized weights (sw):

Table S F1. Summary of the log-hazard ratio from the simulation study with 1,500 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is linear and additive in describing the association between the exposure and the confounder (simulation scenario - I, with weights sw)

	Bias	MSE	SE	SD	Cov.Pr.
Boosted CART	-0.05068	0.00846	0.119	0.0768	0.986
Super learner	-0.01030	0.01135	0.118	0.1061	0.968
Elastic net	-0.00858	0.01470	0.120	0.1209	0.959
Stepwise	0.00918	0.01477	0.123	0.1212	0.959
Bayesian logistic	0.00598	0.01546	0.122	0.1242	0.945
Logistic	0.00728	0.01552	0.122	0.1244	0.945
Pruned CART	-0.09985	0.02458	0.115	0.1209	0.840
CART	-0.16008	0.03841	0.111	0.1131	0.694
Bagged CART	-0.17591	0.04080	0.109	0.0993	0.662
Random Forest	-0.14608	0.23601	0.423	0.4633	0.950
SVM	-0.23728	0.67414	0.157	0.7860	0.799

Table S F2. Summary of the log-hazard ratio from the simulation study with 1,500 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is linear but non-additive in describing the association between the exposure and the confounder (simulation scenario - II, with weights sw).

	Bias	MSE	SE	SD	Cov.Pr.
Boosted CART	0.0561	0.0157	0.139	0.112	0.974
Super learner	0.0254	0.0210	0.142	0.143	0.943
Pruned CART	-0.0695	0.0359	0.139	0.176	0.855
Stepwise	-0.0108	0.0367	0.165	0.191	0.943
CART	-0.1472	0.0383	0.128	0.129	0.819
Bagged CART	-0.1664	0.0428	0.126	0.123	0.753
Elastic net	0.3309	0.1385	0.166	0.170	0.458
Bayesian logistic	0.3631	0.1642	0.173	0.180	0.405
Logistic	0.3682	0.1685	0.174	0.181	0.392
Random Forest	-0.3152	0.5364	0.510	0.661	0.934
SVM	0.2054	0.9616	0.177	0.959	0.505

Table S F3. Summary of the log-hazard ratio from the simulation study with 1,500 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is additive but non-linear in describing the association between the exposure and the confounder (simulation scenario - III, with weights sw).

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	-0.0132	0.0229	0.162	0.151	0.986
Elastic net	0.0735	0.0304	0.158	0.158	0.928
Bayesian logistic	0.0963	0.0346	0.158	0.159	0.905
Logistic	0.0967	0.0347	0.158	0.159	0.905
Boosted CART	-0.0643	0.0454	0.159	0.203	0.964
Bagged CART	-0.3026	0.1215	0.140	0.173	0.419
Pruned CART	0.3098	0.2144	0.250	0.344	0.703
CART	-0.2597	0.2382	0.207	0.413	0.563
Stepwise	0.3491	0.4243	0.296	0.550	0.689
Random Forest	-0.2104	0.7196	0.645	0.822	0.896
SVM	-0.3055	1.2931	0.261	1.095	0.624

Table S F4. Summary of the log-hazard ratio from the simulation study with 1,500 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is non-linear and non-additive in describing the association between the exposure and the confounder (simulation scenario - IV, with weights sw).

	Bias	MSE	SE	SD	Cov.Pr.
Boosted CART	0.1598	0.0446	0.190	0.170	0.977
Super learner	0.2909	0.1388	0.268	0.282	0.932
Elastic net	0.2507	0.1443	0.231	0.240	0.705
Bayesian logistic	0.2756	0.1603	0.235	0.246	0.682
Logistic	0.2766	0.1610	0.235	0.246	0.682
Bagged CART	-0.4426	0.2371	0.150	0.186	0.205
CART	-0.2771	0.3011	0.232	0.507	0.591
Pruned CART	0.3965	0.3554	0.283	0.416	0.545
SVM	-0.0945	1.0882	0.237	1.003	0.167
Stepwise	0.8940	1.2130	0.437	0.698	0.591
Random Forest	-0.3952	1.7614	0.669	1.080	0.886

G. Cumulative Treatment Effect

Tables SG1-SG4 shows the summaries of MSCM results when the hazard function depends on the cumulative treatment exposure (with stabilized weights sw):

Table S G1. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is linear and additive in describing the association between the exposure and the confounder (simulation scenario - V, with weights sw) and the hazard function depends on the cumulative treatment exposure.

	Bias	MSE	SE	SD	Cov.Pr.
Boosted CART	-0.0170	0.0154	0.129	0.123	0.961
Super learner	-0.0348	0.0193	0.129	0.134	0.946
Elastic net	-0.0132	0.0215	0.130	0.146	0.944
Bagged CART	-0.0640	0.0216	0.135	0.132	0.952
Bayesian logistic	-0.0177	0.0262	0.135	0.161	0.932
Logistic	-0.0169	0.0284	0.137	0.168	0.932
CART	0.0315	0.0302	0.148	0.171	0.896
Pruned CART	0.0577	0.0358	0.149	0.180	0.871
Stepwise	0.0074	0.0450	0.139	0.212	0.909
SVM	0.2544	0.3604	0.132	0.544	0.596
Random Forest	-0.8305	1.2149	0.303	0.725	0.254

Table S G2. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is linear but non-additive in describing the association between the exposure and the confounder (simulation scenario - VI, with weights sw) and the hazard function depends on the cumulative treatment exposure.

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	-0.0737	0.0161	0.138	0.140	0.952
Boosted CART	-0.0490	0.0170	0.141	0.116	0.905
Elastic net	-0.0361	0.0170	0.130	0.128	0.952
Bayesian logistic	-0.0415	0.0171	0.132	0.131	0.952
Logistic	-0.0466	0.0178	0.133	0.133	0.952
SVM	-0.0476	0.0234	0.133	0.405	0.882
Bagged CART	-0.1027	0.0283	0.140	0.129	0.905
CART	-0.0988	0.0334	0.149	0.157	0.857
Pruned CART	-0.0642	0.0392	0.145	0.161	0.905
Stepwise	0.2297	0.2431	0.199	0.478	0.857
Random Forest	-0.3636	2.9064	0.514	1.660	0.667

Table S G3. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is additive but non-linear in describing the association between the exposure and the confounder (simulation scenario - VII, with weights sw) the hazard function depends on the cumulative treatment exposure.

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	-0.250	0.102	0.184	0.200	0.6640
Boosted CART	-0.279	0.112	0.174	0.183	0.5780
Elastic net	-0.300	0.133	0.165	0.207	0.4860
Bayesian logistic	-0.299	0.137	0.168	0.217	0.4840
Logistic	-0.305	0.142	0.169	0.221	0.4660
Stepwise	0.106	0.172	0.211	0.401	0.7334
Bagged CART	-0.455	0.248	0.182	0.201	0.2720
Pruned CART	-0.497	0.288	0.172	0.202	0.1860
CART	-0.519	0.316	0.188	0.216	0.2470
SVM	-0.558	0.478	0.158	0.409	0.0892
Random Forest	-1.139	4.476	0.567	1.783	0.6540

Table S G4. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is non-linear and non-additive in describing the association between the exposure and the confounder (simulation scenario - VIII, with weights sw)and the hazard function depends on the cumulative treatment exposure.

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	-0.260	0.110	0.190	0.205	0.6520
Boosted CART	-0.298	0.119	0.178	0.173	0.5500
Elastic net	-0.326	0.147	0.169	0.201	0.4450
Bayesian logistic	-0.325	0.151	0.172	0.213	0.4360
Logistic	-0.330	0.156	0.173	0.218	0.4270
Stepwise	0.101	0.178	0.219	0.410	0.7515
Bagged CART	-0.476	0.264	0.186	0.194	0.2540
Pruned CART	-0.507	0.297	0.176	0.199	0.1820
CART	-0.531	0.326	0.191	0.208	0.2360
SVM	-0.577	0.512	0.161	0.423	0.0917
Random Forest	-1.140	4.445	0.579	1.773	0.6660

H. Mutiple Time-dependent Confounders

Tables SH1-SH4 shows the summaries of MSCM results when the hazard function depends on the cumulative treatment exposure (with stabilized weights sw):

Table S H1. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is linear and additive in describing the association between the exposure and the confounder (simulation scenario - IX, with weights sw) and there exists two time-dependent confounders that affect future treatment status.

	Bias	MSE	SE	SD	Cov.Pr.
Boosted CART	-0.0813	0.0908	0.291	0.290	0.958
Bagged CART	-0.1091	0.1056	0.295	0.306	0.934
Super learner	-0.1089	0.1414	0.305	0.360	0.914
Elastic net	-0.0814	0.1793	0.314	0.416	0.897
CART	-0.1166	0.1838	0.333	0.413	0.879
Bayesian logistic	-0.0908	0.1869	0.319	0.423	0.895
Logistic	-0.0940	0.1993	0.324	0.436	0.891
Pruned CART	-0.1611	0.2053	0.331	0.423	0.860
Stepwise	-0.1042	0.4332	0.352	0.650	0.816
Random Forest	-0.4041	3.5317	0.909	1.835	0.662
SVM	-0.2290	22.9746	0.269	4.788	0.694

Table S H2. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is linear but non-additive in describing the association between the exposure and the confounder (simulation scenario - X, with weights sw) and there exists two time-dependent confounders that affect future treatment status.

	Bias	MSE	SE	SD	Cov.Pr.
Boosted CART	-0.131	0.104	0.304	0.294	0.949
Bagged CART	-0.135	0.109	0.297	0.301	0.942
Super learner	-0.140	0.135	0.305	0.341	0.922
Elastic net	-0.141	0.156	0.305	0.369	0.903
Bayesian logistic	-0.135	0.164	0.311	0.382	0.900
Logistic	-0.131	0.173	0.316	0.396	0.897
CART	-0.133	0.184	0.339	0.408	0.901
Pruned CART	-0.177	0.212	0.339	0.425	0.876
SVM	-0.388	0.393	0.278	0.492	0.767
Stepwise	-0.117	0.403	0.357	0.624	0.856
Random Forest	-0.672	3.340	0.870	1.700	0.658

Table S H3. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is additive but non-linear in describing the association between the exposure and the confounder (simulation scenario - XI, with weights sw) and there exists two time-dependent confounders that affect future treatment status.

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	0.0165	0.0922	0.309	0.303	0.964
Boosted CART	-0.0665	0.1062	0.324	0.319	0.955
Bagged CART	-0.0684	0.1130	0.328	0.329	0.941
CART	0.1078	0.2892	0.413	0.527	0.865
Elastic net	-0.5065	0.3559	0.274	0.315	0.496
Bayesian logistic	-0.4995	0.3614	0.277	0.334	0.497
Logistic	-0.4973	0.3636	0.278	0.341	0.491
Pruned CART	0.1413	0.3859	0.450	0.605	0.832
Stepwise	0.4364	0.5019	0.341	0.558	0.721
SVM	-0.4240	1.7600	0.327	1.257	0.645
Random Forest	-0.4300	2.6455	0.842	1.569	0.698

Table S H4. Summary of the log-hazard ratio from the simulation study with 250 subjects, each with up to 10 visits (1,000 Monte Carlo dataset) where the treatment selection model is non-linear and non-additive in describing the association between the exposure and the confounder (simulation scenario - XII, with weights sw) and there exists two time-dependent confounders that affect future treatment status.

	Bias	MSE	SE	SD	Cov.Pr.
Super learner	-0.0387	0.125	0.335	0.351	0.949
Boosted CART	-0.1978	0.170	0.353	0.362	0.921
Bagged CART	-0.2158	0.203	0.358	0.396	0.879
CART	-0.0638	0.340	0.434	0.579	0.848
Elastic net	-0.4931	0.402	0.300	0.398	0.548
Bayesian logistic	-0.4687	0.406	0.303	0.432	0.548
Logistic	-0.4619	0.407	0.303	0.440	0.550
Pruned CART	-0.0998	0.413	0.449	0.635	0.824
Stepwise	0.4083	0.566	0.374	0.632	0.736
SVM	-0.4558	1.350	0.332	1.069	0.583
Random Forest	-0.3598	59.520	0.885	7.707	0.663

I. Simulation Graphs

Figures SI1-SI12 shows the trends of bias, MSE and coverage probabilities of MSCM estimates when weights are progressively truncated at higher percentiles in each simulation scenario.

Simulation - I

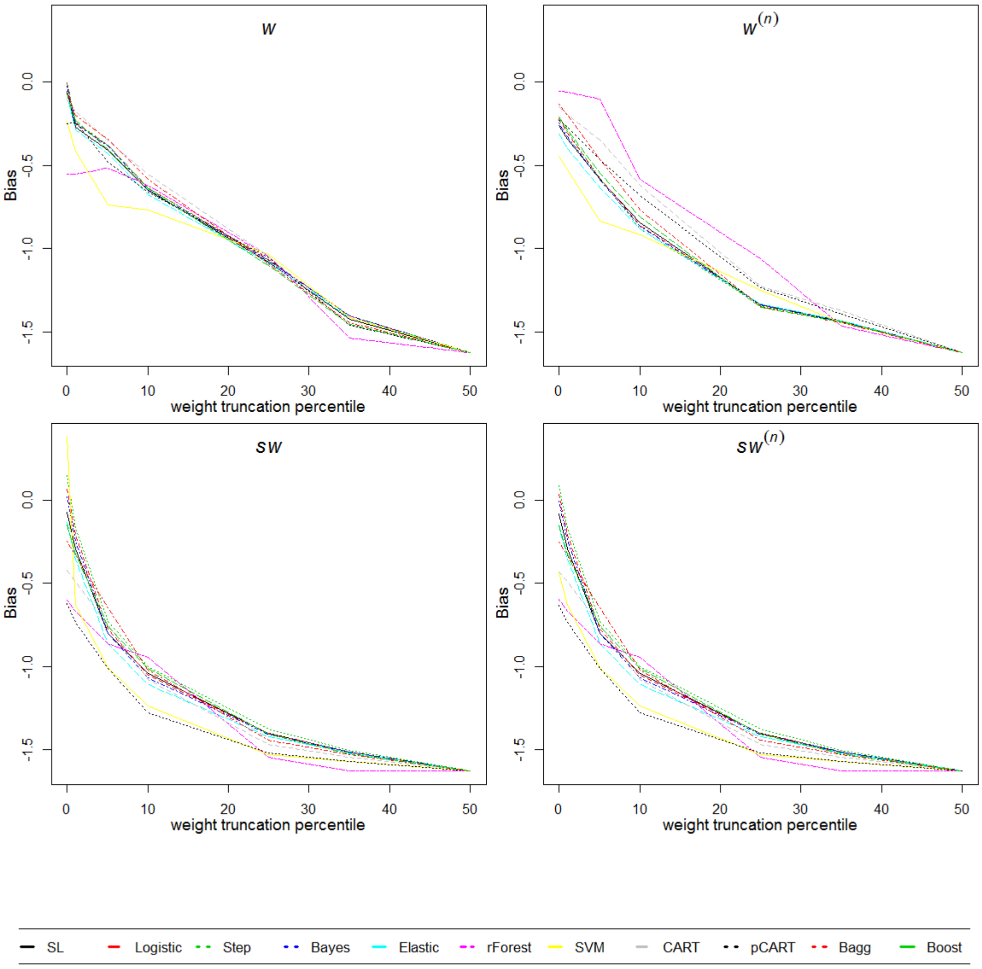

Figure S I1. Bias in MSCM estimate $\hat{\psi}_1$ when the weights are progressively truncated in a simulation study of $1,000$ datasets with 250 subjects observed at most 10 times (Simulation - I).

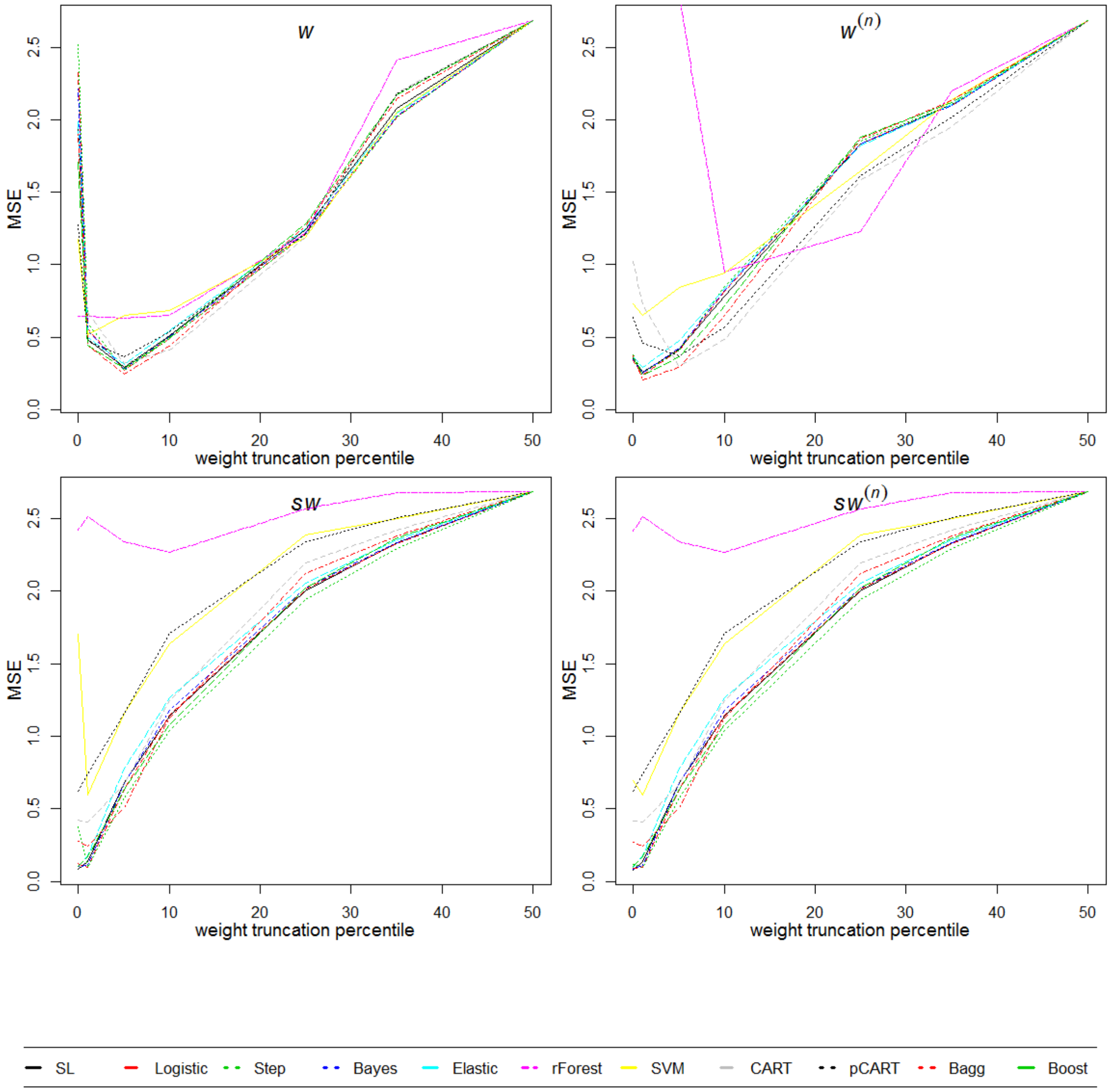

Figure S I2. Mean squared error in MSCM estimate $\hat{\psi}_1$ when the weights are progressively truncated in a simulation study of $1,000$ datasets with 250 subjects observed at most 10 times (Simulation - I).

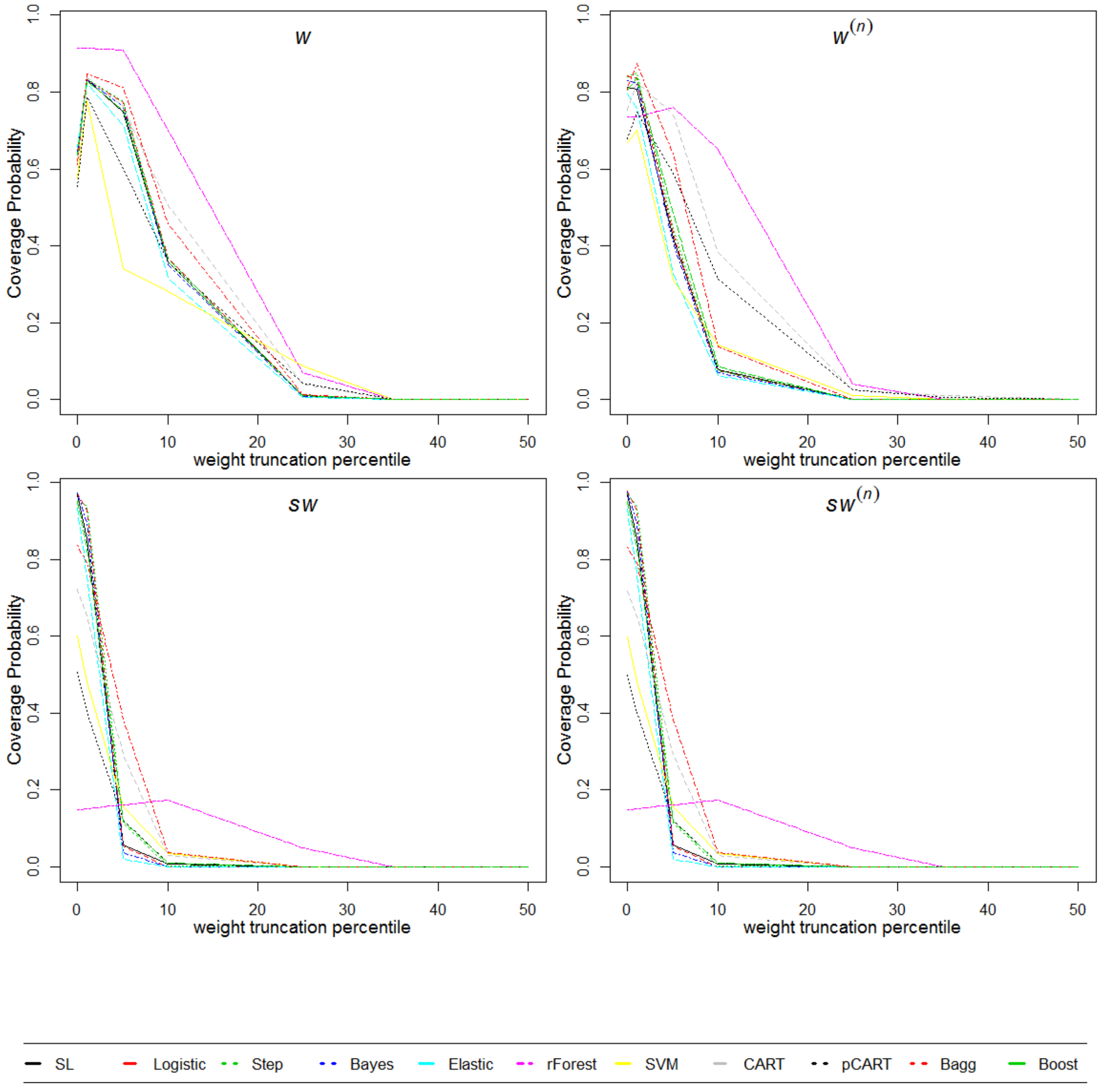

Figure S I3. The coverage probability of model-based nominal 95% confidence intervals based on the MSCM estimate $\hat{\psi}_1$ when the weights are progressively truncated in a simulation study of $1,000$ datasets with 250 subjects observed at most 10 times (Simulation - I).

Simulation - II

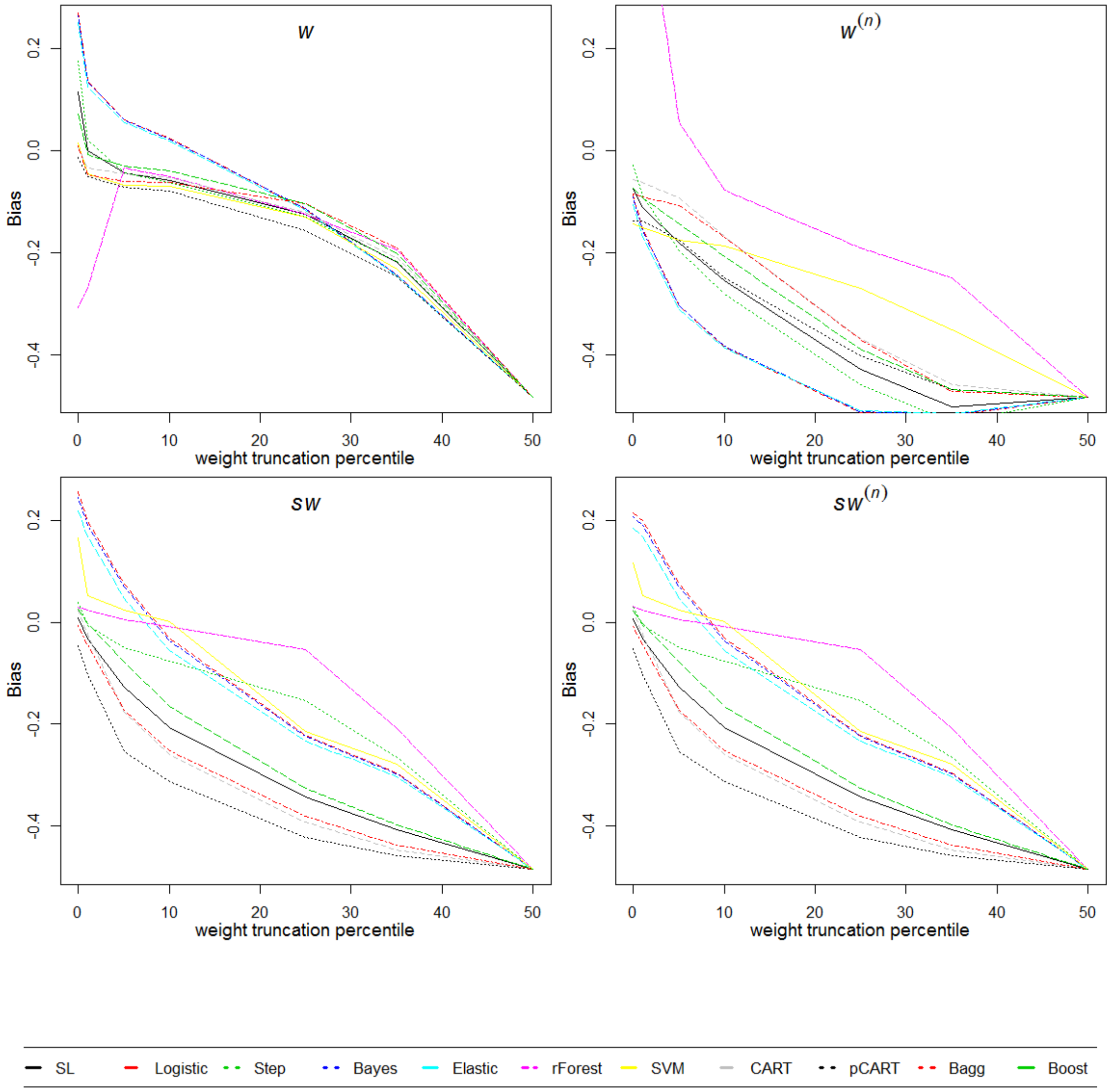

Figure S I4. Bias in MSCM estimate $\hat{\psi}_1$ when the weights are progressively truncated in a simulation study of $1,000$ datasets with 250 subjects observed at most 10 times (Simulation - II).

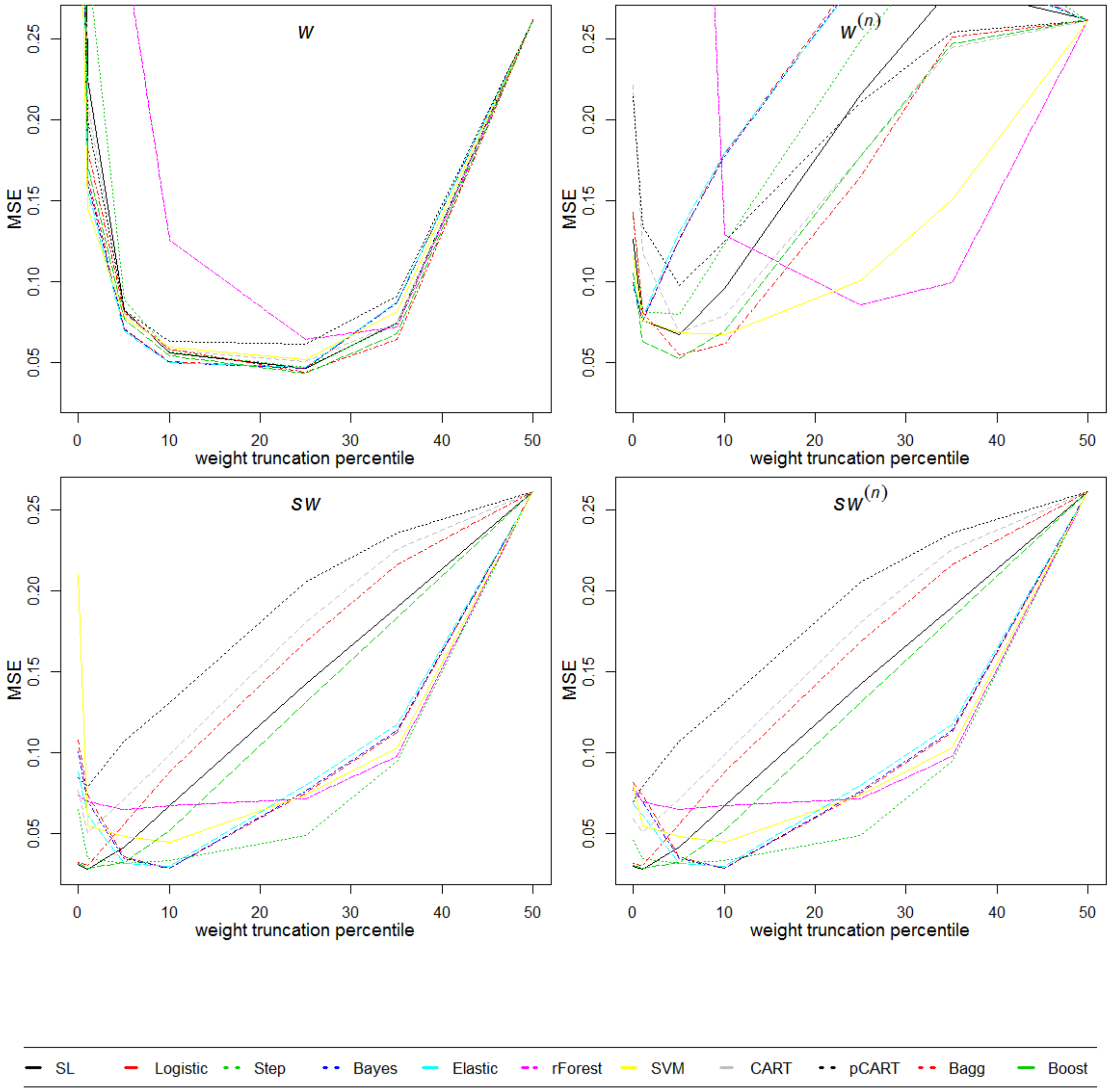

Figure S I5. Mean squared error in MSCM estimate $\hat{\psi}_1$ when the weights are progressively truncated in a simulation study of $1,000$ datasets with 250 subjects observed at most 10 times (Simulation - II).

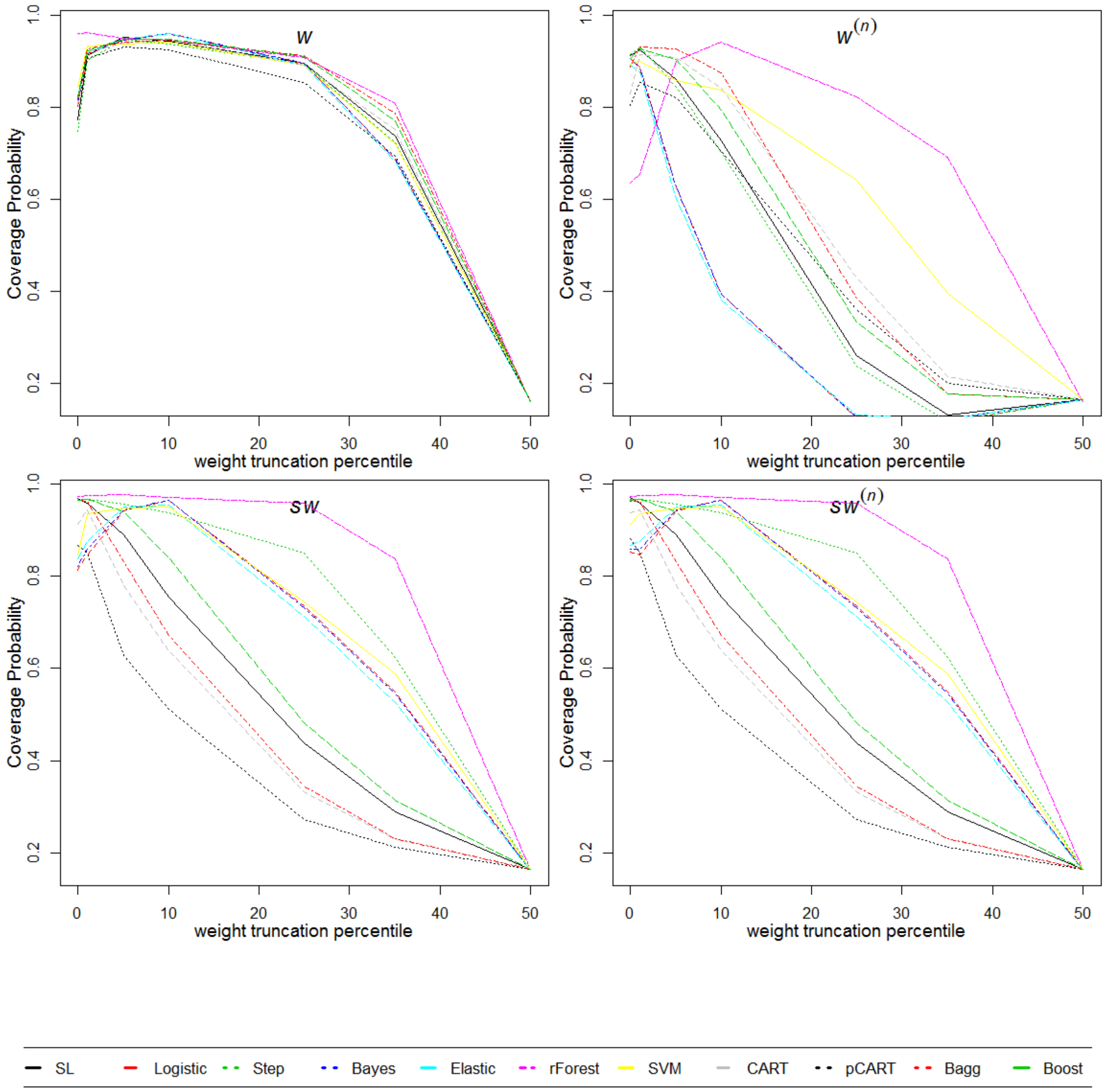

Figure S I6. The coverage probability of model-based nominal 95% confidence intervals based on the MSCM estimate $\hat{\psi}_1$ when the weights are progressively truncated in a simulation study of 1,000 datasets with 250 subjects observed at most 10 times (Simulation - II).

Simulation - III

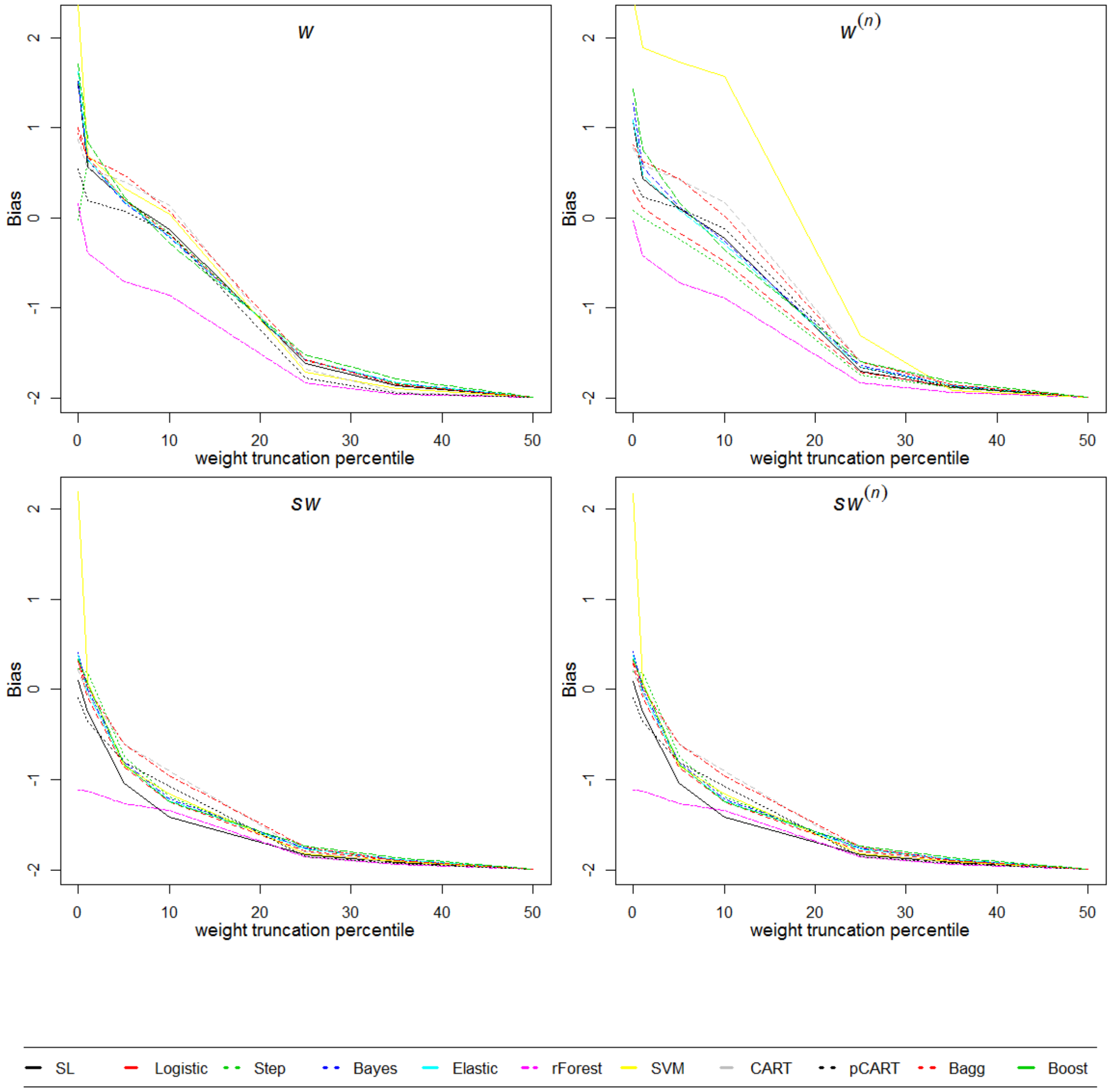

Figure S I7. Bias in MSCM estimate $\hat{\psi}_1$ when the weights are progressively truncated in a simulation study of $1,000$ datasets with 250 subjects observed at most 10 times (Simulation - III).

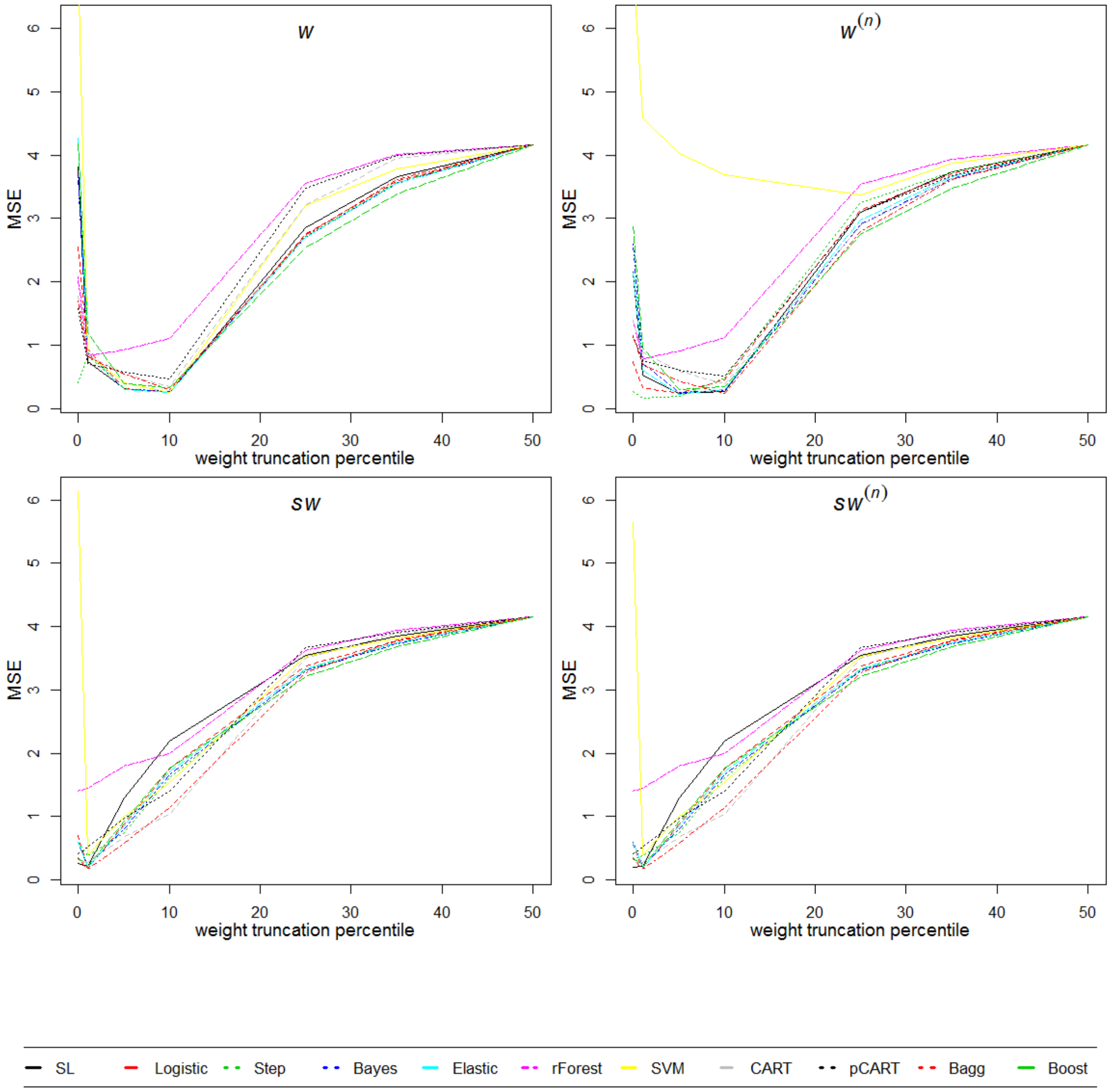

Figure S I8. Mean squared error in MSCM estimate $\hat{\psi}_1$ when the weights are progressively truncated in a simulation study of $1,000$ datasets with 250 subjects observed at most 10 times (Simulation - III).

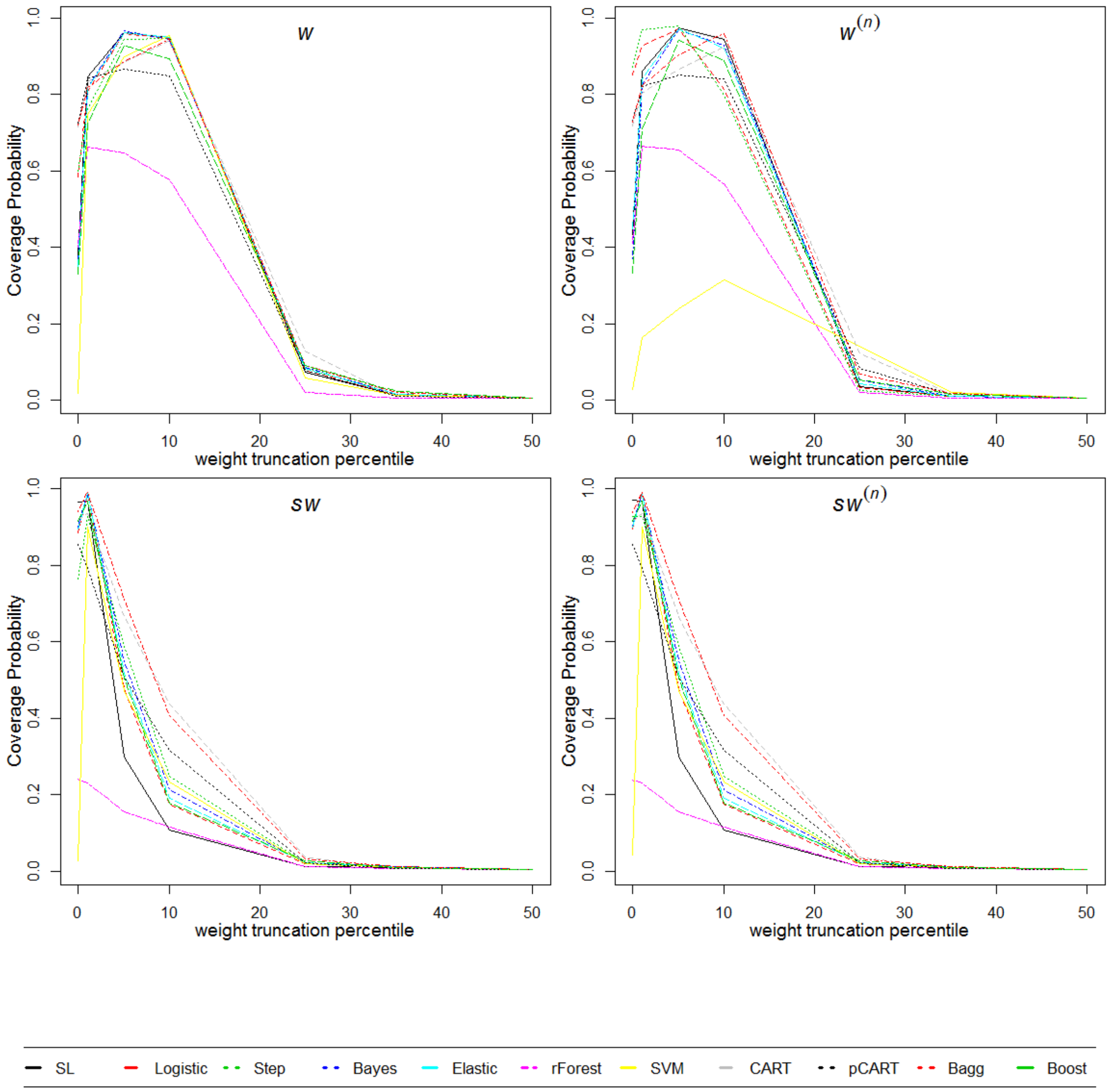

Figure S I9. The coverage probability of model-based nominal 95% confidence intervals based on the MSCM estimate $\hat{\psi}_1$ when the weights are progressively truncated in a simulation study of $1,000$ datasets with 250 subjects observed at most 10 times (Simulation - III).

Simulation - IV

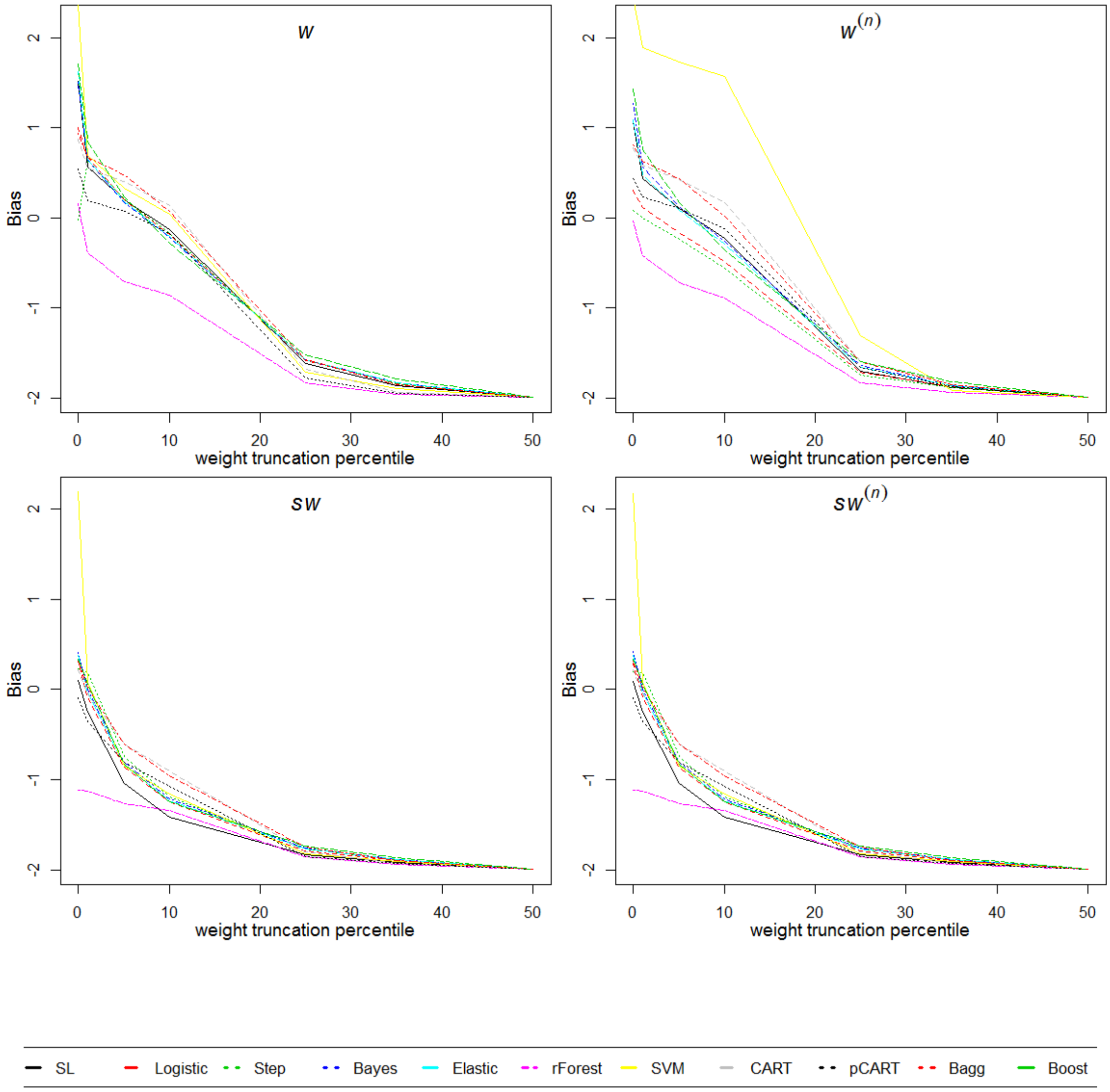

Figure S I10. Bias in MSCM estimate $\hat{\psi}_1$ when the weights are progressively truncated in a simulation study of $1,000$ datasets with 250 subjects observed at most 10 times (Simulation - IV).

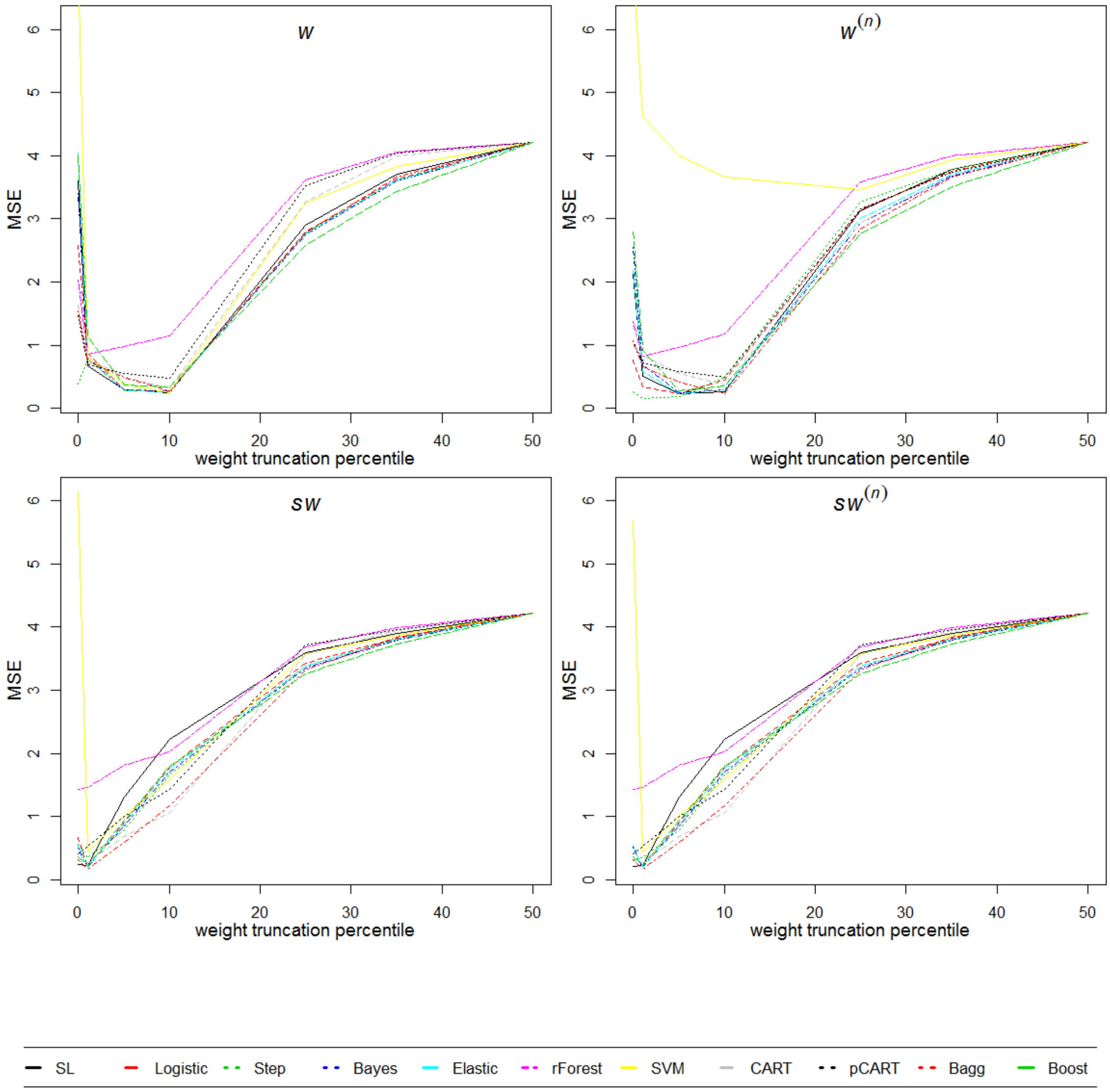

Figure S I11. Mean squared error in MSCM estimate $\hat{\psi}_1$ when the weights are progressively truncated in a simulation study of $1,000$ datasets with 250 subjects observed at most 10 times (Simulation - IV).

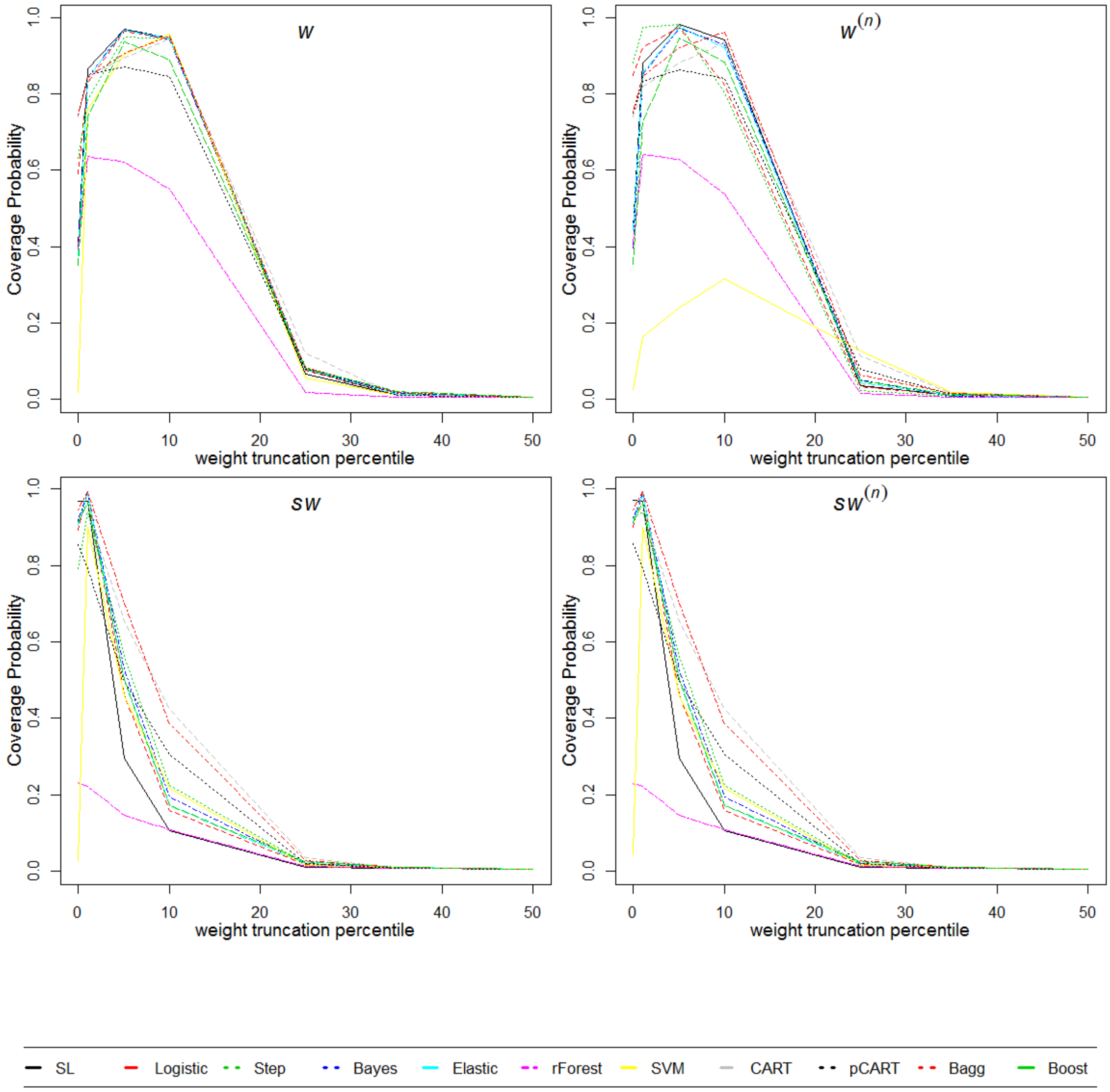

Figure S I12. The coverage probability of model-based nominal 95% confidence intervals based on the MSCM estimate $\hat{\psi}_1$ when the weights are progressively truncated in a simulation study of $1,000$ datasets with 250 subjects observed at most 10 times (Simulation - IV).

J. Summary of the Selected Cohort and Exclusion Criteria

The eligibility for β-IFN treatment in BC was adapted from the provincial government's reimbursement scheme. This criteria used for β-IFN treatment are: patients have to be at least 18 years old, have an Expanded Disability Status Scale (EDSS) score of 6.5 or below (i.e., able to walk 20 meters without resting with constant bilateral support) and have definite MS with a relapsing-onset course.

Table S J1. Characteristics of the selected cohort of patients with relapsing-onset multiple sclerosis (MS), British Columbia, Canada (1995-2008).

Baseline characteristics	Ever-β-IFN exposed	Never-β-IFN exposed
Number	868	829
Women, n (%)	660 (76.0)	637 (76.8)
Disease duration, average (SD)	5.8 (6.6)	8.3 (8.5)
Age, average (SD)	38.1 (9.2)	41.3 (10.0)
EDSS score, median (range)	2.0 (0-6.5)	2.0 (0-6.5)
Relapse rate / year,[†] median (IQR)	0.5 (0-1.2)	0.5 (0-1.0)
Active follow-up time,[‡] average (SD)	5.2 (2.8)	4.5 (2.9)

[†] Over the 2 years prior to baseline.

[‡] First to last EDSS measurement, measured in years.

$2,671$ patients met the eligibility criteria to receive β-IFN treatment between July 1995 and December 2004. Of these, patients who were exposed to a non-β-IFN immunomodulatory drug, a cytotoxic immunosuppressant for MS ($n = 172$), or an MS clinical trial ($n = 21$) prior to baseline were excluded from the analysis. If the exposure occurred after baseline, data were censored at the start of the exposure to the non-β-IFN treatment. Further exclusion criteria included unknown MS onset date ($n = 10$), insufficient EDSS measurements ($n = 436$), reaching of the outcome ($n = 218$) or the secondary progressive stage before the eligibility date ($n = 217$). Some patients met multiple exclusion criteria. As a result, $1,697$ patients were selected. A summary of their characteristics are reported in Table SJ1.

K. Additional Analysis of Multiple Sclerosis Data

Table S K1. The impact of truncation of the sw (generated via super learner) on the estimated causal effect of β-IFN on reaching sustained EDSS 6 for BC MS patients (1995-2008).

Truncation	Estimated weights		Treatment effect estimate		
percentiles	Mean (log-SD)	Min-Max	HR	SE[†]	95% CI[†]
None	1.056 (-0.771)	0.392 - 2.379	1.349	0.316	0.853 - 2.134
(1, 99)	1.056 (-0.773)	0.443 - 2.030	1.278	0.278	0.834 - 1.957
(5, 95)	1.055 (-0.782)	0.469 - 1.965	1.187	0.241	0.797 - 1.767
(10, 90)	1.051 (-0.808)	0.486 - 1.898	1.215	0.236	0.830 - 1.778
(25, 75)	0.990 (-1.404)	0.693 - 1.310	1.234	0.223	0.866 - 1.760
(35, 65)	0.973 (-2.000)	0.818 - 1.124	1.253	0.222	0.886 - 1.772
Median[‡]	0.995 (-Inf)	0.995 - 0.995	1.288	0.225	0.914 - 1.815

log-SD, logarithmic transformation of standard deviation; Min, minimum; Max, maximum; CI, confidence interval; HR, Hazard ratio; SE, standard error.

[†] Based on robust standard error.

[‡] Baseline-adjusted analysis.

-