Optimal probability weights for estimating causal effects of time-varying treatments with marginal structural Cox models

Michele Santacatterina*, Celia García-Pareja
Rino Bellocco, Anders Sönnerborg, Anna Mia Ekström
and Matteo Bottai
Karolinska Institutet, Stockholm, Sweden 17177

*The authors are grateful to the KID grant program at Karolinska Institutet and an ALF grant, Stockholm, Sweden, for the provided support. The authors also thank Dr. Erica E.M. Moodie and Dr. Xiao Yongling for their help on the simulation setup.

Abstract

Marginal structural Cox models have been used to estimate the causal effect of a time-varying treatment on a survival outcome in the presence of time-dependent confounders. These methods rely on the positivity assumption, which states that the propensity scores are bounded away from zero and one. Practical violations of this assumption are common in longitudinal studies, resulting in extreme weights that may yield erroneous inferences. Truncation, which consists of replacing outlying weights with less extreme ones, is the most common approach to control for extreme weights to-date. While truncation reduces the variability in the weights and the consequent sampling variability of the estimator, it can also introduce bias. Instead of truncated weights, we propose using optimal probability weights, defined as those that have a specified variance and the smallest Euclidean distance from the original, untruncated weights. The set of optimal weights is obtained by solving a constrained quadratic optimization problem. The proposed weights are evaluated in a simulation study and applied to the assessment of the effect of treatment on time to death among people in Sweden who live with human immunodeficiency virus and inject drugs.

Keywords: Causal inference, longitudinal data, positivity assumption, probability weights, survival analysis.

1 Introduction

Marginal structural Cox models (MSCM) (Robins et al., 2000; Hernán et al., 2000) have been used to estimate the causal effect of a time-varying treatment on a survival outcome with observational data. The increasing popularity of MSCM derives from their ability to handle time-dependent confounders, which are confounders that are affected by previous treatments and affect future ones (Daniel et al., 2013). For example, the HIV-Causal Collaboration (HIV-Causal Collaboration, 2011) used MSCM to evaluate the optimal timing of human immunodeficiency virus (HIV) treatment initiation on time to death, where CD4 cell count was both a predictor of treatment initiation and survival, as well as being itself influenced by prior treatment. Standard procedures, such as regression adjustment or matching, fail to control for time-dependent confounding, thus introducing post-treatment bias (Blackwell, 2013; Robins, 2000). MSCM are estimated via inverse probability of treatment weighting (IPTW) (Hernan and Robins, 2010), which controls for time-dependent confounding by creating a hypothetical population where time-dependent and time-invariant confounders are balanced over time (Cole and Hernán, 2008). These weights are constructed as the inverse of the product of the probabilities of being assigned to the treatment conditional on covariates and treatment history, i.e. the propensity scores (Rosenbaum and Rubin, 1983) estimated separately at each time point (Cole and Hernán, 2008). Despite their theoretical appeal and their wide range of applications, IPTW-based methods are sensitive to violations of the positivity assumption, also referred to as the experimental treatment assignment assumption (Imbens and Rubin, 2015). This states that the propensity score of each unit under study is bounded away from zero and one. Positivity is practically violated when subjects in specific strata of the population under study have a low probability of receiving the treatment, leading to extreme weights, erroneous

inferences, and low precision (Robins et al., 1995; Scharfstein et al., 1999; Robins et al., 2007; Kang and Schafer, 2007).

Several methods have been developed to alleviate the problems caused by extreme weights when considering one single time point (Santacatterina and Bottai, 2017; Zubizarreta, 2015; Hainmueller, 2012; Athey et al., 2016). With longitudinal data, truncation, which consists of replacing outlying weights with less extreme ones, remains the most popular solution to this problem (Cole et al., 2005). However, while truncation reduces the variability of the weights, thus increasing inferential precision, it can also introduce considerable bias. Ad-hoc and empirical criteria have been proposed to choose the truncation threshold. Under the assumption that the MSCM estimates are unbiased, Cole et al. Cole et al. (2005) suggested choosing the truncation level by progressively truncating the weights until a trade-off between bias and variance is found. Xiao et al. Xiao et al. (2013) compared different truncation levels for MSCM, and proposed a data-adaptive approach to select the best level of truncation that minimizes the mean squared error. The authors showed an improvement in the MSCM estimates when truncating the weights at high percentiles of their distribution. Methods other than truncation have been proposed, including history-restricted MSCM Neugebauer et al. (2007) where information on a restricted portion of the treatment history is used to estimate the causal effects, trimming Stürmer et al. (2010) where observations that violate the positivity assumption are excluded, and G-computation Robins (2000), a non-IPTW-based method.

The purpose of this paper is to introduce optimal probability weights Santacatterina and Bottai (2017) (OPW) to the estimation of the causal effect of a time-varying treatment with longitudinal data when the positivity assumption is practically violated. OPW are the solution to a constrained quadratic optimization problem, which finds the closest set of weights to the original, untruncated weights while controlling the precision of the resulting

weighted estimator. Differently from Santacatterina and Bottai Santacatterina and Bottai (2017), this paper focuses on repeated observations. In addition, the constraint is placed on the variance of the weights instead of the variance of the weighted estimator. This formulation of the optimization problem is novel and has two main advantages: (1) it is quadratic and convex and therefore admits a unique solution; and (2) it is independent of both the chosen estimator for the causal parameter of interest and that for its standard error.

The following section briefly reviews MSCM. Section 3 introduces the quadratic problem used to obtain the set of optimal probability weights, describes their properties, and discusses the choice of the parameter that controls precision. Section 4 shows the results of a simulation study. Section 5 presents an application of the optimal probability weights to the evaluation of the effect of HIV treatment initiation on time to death among people in Sweden who inject drugs. Final conclusions are given in Section 6.

2 Marginal structural Cox models

We consider a longitudinal study where n units are observed at regular time intervals $k = 1, \ldots, K$ (e.g. every 3 months). For each unit $i = 1, \ldots, n$, we denote by T_i the observed follow-up time, and by V_i the vector of baseline covariates. For each unit i at time t, we denote by $A_i^{(t)}$ the binary time-varying treatment variable, where $A_i^{(t)} = 0$ means not being treated at time t, and $A_i^{(t)} = 1$ means being treated at time t, and by $X_i^{(t)}$ the time-dependent covariates. We assume that the treatment $A_i^{(t)}$ and the covariates $X_i^{(t)}$ do not change between two time intervals $(k, k+1)$. We denote by $\overline{A}_i^{(t)}$ the treatment history up to time t and, $\overline{X}_i^{(t)}$ the covariates history up to time t, i.e. the time-dependent confounders' history. We define $Y_i^{(t)}$ the event at time t, which equals 1 if the subject i had

the event at time t, and 0 otherwise. Finally, we denote by $T_{\overline{a}^{(t)}}$ the counterfactual failure time, had the subject followed the treatment history $\overline{a}^{(t)} = \{a^{(t)}; 0 \leq t < \infty\}$. For each $\overline{a}^{(t)}$, we define the MSCM as follows,

$$\lambda_{T_{\overline{a}^{(t)}}}(t|V) = \lambda_{\overline{0}^{(t)}}(t)\exp(\beta_1\gamma(\overline{a}^{(t)}) + \beta_2 V) \tag{2.1}$$

where $\lambda_{T_{\overline{a}^{(t)}}}(t|V)$ is the hazard at time t given baseline covariates V had, contrary to fact, the subject followed the treatment history $\overline{a}^{(t)}$, $\lambda_{\overline{0}^{(t)}}(t)$ is the baseline hazard at time t for a never-treated subject $\overline{a}^{(t)} = \overline{0}^{(t)}$ with $V = 0$, $\gamma(\cdot)$ is a known function for the treatment history, and β_1 is the causal parameter of interest. Under the assumptions of positivity, consistency, no unmeasured confounders, and correct specification of the models, the causal parameter β_1 can be consistently estimated using IPTW Hernán et al. (2000); Cole and Hernán (2008). The stabilized version of the inverse probability of treatment weights can be obtained as follows Hernán et al. (2000)

$$w_*^{(t)} = \prod_{k=1}^{m(t)} \frac{Pr(A^{(k)} = a^{(k)}|\overline{A}^{(k-1)} = \overline{a}^{(k-1)}, V = v)}{Pr(A^{(k)} = a^{(k)}|\overline{A}^{(k-1)} = \overline{a}^{(k-1)}, \overline{X}^{(k)} = \overline{x}^{(k)}, V = v)} \tag{2.2}$$

where $m(t)$ is the number of visits up to time t. When informative censoring is present, under all the aforementioned assumptions, and with the additional assumption of no unmeasured informative censoring, the causal parameter β_1 can be consistently estimated using weights obtained by the product of inverse probability of treatment and inverse probability of censoring weights Hernán et al. (2001). The set of inverse probability of censoring weights is computed similarly to that of equation (2.2). Parametric models, such as logistic regression, are commonly used to estimate $w_*^{(t)}$, along with machine learning methods, such as support vector machines and classification and regression trees Karim et al. (2017). Throughout this paper, we refer to $\hat{w}_*^{(t)}$, the estimated weights used to control for

time-dependent confounding, as the set of *target* weights.

3 Optimal probability weights

When the positivity assumption is practically violated, the estimated set of target weights $\hat{w}_*^{(t)}$ may contain outliers, which may yield low precision and erroneous inferences on the causal parameter β_1. As suggested by Santacatterina and Bottai Santacatterina and Bottai (2017), rather than truncating, we propose to obtain weights $\hat{w}_o^{(t)}$ that are the closest to $\hat{w}_*^{(t)}$ with respect to the Euclidean norm, while constraining the variance of the weights $\hat{w}_o^{(t)}$ to be less or equal to a specified level ξ. The resulting quadratic optimization problem can be formulated as follows.

$$\underset{w_o^{(t)} \in \mathbb{R}^{n \times t}}{\text{minimize}} \quad \|w_o^{(t)} - \hat{w}_*^{(t)}\|_2 \tag{3.1}$$

$$\text{subject to} \quad \|w_o^{(t)} - \overline{w}_o^{(t)}\|_2^2 \leq \xi \tag{3.2}$$

$$w_o^{(t)} \geq 0 \tag{3.3}$$

where $\overline{w}_o^{(t)}$ is the mean of the weights $w_o^{(t)}$. Constraint (3.2) controls the variance of the weights, and therefore the precision of the resulting weighted estimator. Constraint (3.3) ensures that the weights are non-negative. We refer to $\hat{w}_o^{(t)}$, solution to the problem (3.1)-(3.3), as the set of *optimal* probability weights (OPW). Santacatterina and Bottai Santacatterina and Bottai (2017) showed that the weighted estimator that uses optimal weights $\hat{w}_o^{(t)}$ is consistent. They also showed that if the weighted estimator that uses target weights $\hat{w}_*^{(t)}$ is unbiased, minimizing the distance between $\hat{w}_o^{(t)}$ and $\hat{w}_*^{(t)}$ is equivalent to minimizing the bias of the weighted estimator that uses $\hat{w}_o^{(t)}$. They concluded that high precision could be reached with a low increase in bias, in all the scenarios considered in their

simulations. Finally, the objective function and the constraint in the proposed quadratic problem (3.1)-(3.3) are convex, therefore admitting a unique solution.

3.1 On the choice of ξ

The solution to the quadratic problem (3.1)-(3.3) depends on the constant ξ, which controls the variance of the weights and consequently the precision of the estimates. We suggest choosing ξ in function of the aims of the study. The following are some practical guidelines.

1. *Variance of weights obtained by truncation.* Xiao et al. Xiao et al. (2013) suggested that truncation at high percentiles, such as the 99th or the 99.5th percentile of the distribution of the target weights improves the IPTW estimators. Therefore, one can truncate the target weights at high percentiles, compute their variance, and set ξ equal to the variance of the obtained truncated weights. In Section 4.2, we show how the MSCM that uses OPW performs better, in terms of mean squared error, than that using truncated weights especially when the weights are truncated at high percentiles.

2. *Evaluation of the Lagrange multiplier.* Constraint (3.2) in the quadratic problem (3.1)-(3.3) has an associated Lagrange multiplier, λ_L, which can provide insight on the relationship between the optimal solution and the constraint. Specifically, small values of λ_L suggest that a small decrease in ξ would lead to a small increase in the optimal value of the objective function (3.1). Large values of λ_L suggest that a small decrease in ξ would lead to a large increase in the optimal value of the objective function. Consequently, λ_L may be used to select the level of precision ξ. In Section 4.2 we show how λ_L reflects the behavior of the bias across different levels of precision.

3. *Bias-variance trade-off.* Cole and Hernán Cole and Hernán (2008) suggested using truncation as a means to trade off bias and variance. If the untruncated IPTW estimate, weighted by the set of target weights $\hat{w}_*^{(t)}$, is unbiased for the causal parameter of interest, minimizing the objective function in (3.1) leads to minimizing the bias of the IPTW estimator that uses the set of optimal weights, while controlling the precision of the resulting IPTW estimator. A grid of values for ξ may be used to evaluate the bias-variance trade-off. As in Cole and Hernán Cole and Hernán (2008), an acceptable value for ξ may be selected after investigating the values of the estimated weighted parameter and its estimated standard error against the grid of levels of ξ.

4. *Pre-specified level of precision.* Similarly to sample size and power calculations, the level of ξ may be set to match a pre-specified, desired precision of the resulting MSCM estimates.

5. *Variance of the weights obtained with simplified weights models.* Deep classification trees and logistic regression models with many covariates and higher-order interactions can estimate the set of target weights. This yields nearly unbiased but highly variable estimates of the causal parameters. Simplifying these models by considering, for example, a logistic regression model with only the main effects or a less deep tree, may increase the precision. The value for ξ may be set to be equal to that obtained with the simplified model.

4 Simulations

In this section, we present the setup and results of a simulation study designed to compare OPW, solution to (3.1)-(3.3), and weights truncated at different levels with respect to mean squared error (MSE), bias, and standard error of the MSCM estimator. The study is aimed at mimicking data from a longitudinal study of a hypothetical cohort of HIV-positive patients Xiao et al. (2013), similar to that discussed in Section 5.

4.1 Setup

We randomly generated 1,000 samples, each of which comprised 200 or 1,000 observations using a maximum follow-up time of $K = 10$ biyearly visits. For each interval $k = 1, \ldots, K$, we generated the expected survival time t^* by using the quantile function of an exponential distribution with the interval-specific hazard rate computed from the following model

$$\lambda_{i,k}(t^*|A_i^{(k)}, X_i^{(k)}) = \lambda_0(t^*)\exp(\theta_1 A_i^{(k)} + \theta_2 X_i^{(k)}) \tag{4.1}$$

with $\lambda_0(t^*) = 0.12$, $\theta_1 = \log(0.5)$, $\theta_2 = -0.0016$, $A_i^{(k)} \sim$ Binomial(π), $\pi = (1 + \exp(3.623 - 2.605I[X_i^{(k)} > 500] - 0.022(X_i^{(k)} - 200) + 0.009(X_i^{(k)} - 200)I[X_i^{(k)} > 500] + 0.405A_i^{(k-1)})^{-1}$ for $k \leq 1 \leq K$, $A_i^{(0)} = 0$, $X_i^{(k)} = X_i^{(k-1)} + 70A_i^{(k-1)} + \Delta_i + \varepsilon_i$, $\varepsilon_i \sim$ Normal$(0, 3)$, $\Delta_i \sim$ Uniform$(-80, -5)$ for $k \leq 2 \leq K$, and $X_i^{(1)} = V_i =$ Lognormal(6,1) . We defined the observed follow-up time as $t_i = \min(T_i, C_i, 5)$, where $T_i = 0.5(k-1) + t_i^*(k)$ for $1 \leq k < K$ and $T_i = 5$ for $k \geq K$, and $C_i \sim$ Uniform$(0, 40)$. The true causal parameter of interest, the hazard ratio (HR), was set to be equal to HR = 0.5. A detailed explanation of the data generating process is provided by Xiao et al. Xiao et al. (2010). We also considered two additional scenarios in which the practical positivity assumption was weakly and strongly violated. Specifically, under the weak violation scenario we considered

$\pi = (1 + \exp(1.623 - 0.605I[X_i^{(k)} > 500] - 0.0015(X_i^{(k)} - 200) + 0.405A_i^{(k-1)})^{-1}$, which provided almost uniformly distributed weights, while under the strong violation scenario we considered $\pi = (1 + \exp(4.623 - 2.605I[X_i^{(k)} > 500] - 0.02(X_i^{(k)} - 200) + 0.009(X_i^{(k)} - 200)I[X_i^{(k)} > 500] + 0.405A_i^{(k-1)})^{-1}$, which provided more extreme weights than the original setting aforementioned. In particular, for $n = 1,000$, under the weak violation scenario, the mean of the weights across simulations ranged between 0.9659 and 1.0660, under the scenario provided by Xiao et al. Xiao et al. (2010) between 0.6203 and 18.3, while under the strong violation scenario between 0.3390 and 51.77. We considered the set of stabilized inverse probability weights as the target weights of interest. Truncated weights were obtained by truncating the set of target weights across different quantiles defined as a grid of twenty equally-spaced values between 0.8 and 1. OPW were obtained by solving (3.1)-(3.3) with ξ equal to the variance of the truncated weights for each of the different levels of truncation. In each simulated sample, we estimated the causal parameter of interest by using the following Cox regression model

$$\lambda_{i,k}(t|\overline{A}_i^{(k)}, V_i) = \lambda_0(t)\exp(\beta_1 A_i^{(k)} + \beta_2 A_i^{(k-1)} + \beta_3 V_i) \tag{4.2}$$

weighted by the truncated weights and by the set of OPW. We used a robust estimator of the standard error Austin (2016). We estimated the stabilized inverse probability of treatment weights using the R package ipw van der Wal et al. (2011), and we solved the quadratic problem (3.1)-(3.3) by using the package lpoptr Wächter and Biegler (2005) and the MA57 sparse symmetric system as line-search method HSL (2017). We provide the R code for the simulations as Supporting Information.

4.2 Results

The top-left panels of Figure 1 and Figure 2 show the MSE ratio between the hazard ratio estimated with truncated weights and that estimated with OPW across truncation levels when $n = 200$ and $n = 1,000$, respectively. The optimally weighted MSCM performed better than the truncated MSCM at all truncation levels, especially between the 98th and the 99.5th percentile of the distribution of the target weights. In particular, the value between the 98th and the 99.5th percentile for which the MSE ratio is largest is equal to the 99.5th percentile when $n = 200$ and equal to the 99th percentile when $n = 1,000$. When truncating at lower percentiles, optimally weighted and truncated MSCM performed equally in small samples ($n = 200$), but not in larger samples ($n = 1,000$) where the optimally weighted MSCM showed a substantially smaller MSE. At the lowest truncation levels and with the smaller sample size, the distributions of truncated weights and that of the OPW were almost uniformly distributed, resulting in a similar MSE. In the larger samples, the bias of the truncated MSCM increased with increasing levels of truncation while that of the optimally weighted MSCM remained almost constant. The top-right panels of Figure 1 and Figure 2 show the MSE (solid line), variance (dotted), and bias (dashed) of the estimated hazard ratio that uses OPW across truncation levels. Setting the constant ξ based on high-percentile truncation weights improves the behaviour of the MSCM by introducing small bias but considerably increasing precision. The mean solving time of the algorithm was below 0.22 seconds in the smaller samples and below 1.0 second in the larger samples (bottom-left panels of Figure 1 and Figure 2). The standardized mean Lagrange multiplier associated with constraint (3.2) partially reflected the behaviour of the bias (bottom-right panels of Figure 1 and Figure 2), and it may be used to choose ξ as discussed in Section 3.1. Figure 3 shows scatter-plots and histograms of the mean truncated weights, (X-axis), and the mean OPW, (Y-axis), across simulations when $n = 1,000$ for each of the four

thresholds, 1, 0.99, 0.85 and 0.80. Weights were first scaled to have mean 0 and variance 1 and then log-transformed. In particular, the top-left panel of Figure 3 shows the original untruncated distribution of the weights, which was asymmetric with a long right tail. For the remaining thresholds, 0.99, 0.85 and 0.80, truncated weights showed a wider distribution compared with OPW. For instance, when the threshold was set to be equal to 0.80, the set of OPW ranged between -0.0188 and 0.0504, while that of the truncated weights between -0.1705 and 0.2767. The left panels of Figure 4 show the MSE ratio between the hazard ratio estimated with truncated weights and that estimated with OPW across truncation levels when $n = 1,000$ under the weak and strong violation scenarios. Under weak violation of the positivity assumption we observed no differences between the truncated MSCM and the optimally weighted MSCM. Under the strong violation scenario, however, the optimally weighted MSCM showed a consistently smaller MSE across truncation levels, and a greater precision compared with the scenario presented in Figure 2, i.e., the MSE ratios were larger under the strong scenario. The right panels of Figure 4 show the MSE (solid line), variance (dotted), and bias (dashed) of the estimated hazard ratio that uses OPW across truncation levels. Under weak violation, no differences were seen across truncation levels, while under strong violation, similarly to the scenario presented in Figure 2, the constant ξ based on high-percentile truncation weights improved the behaviour of the MSCM by introducing small bias but significantly increasing precision. We conclude that OPW were more narrowly distributed, thus leading to more precise inferences, than truncated weights across all considered thresholds, and that OPW outperformed truncated weights across both sample sizes and different scenarios of practical positivity violation, especially under strong violations of the practical positivity assumption.

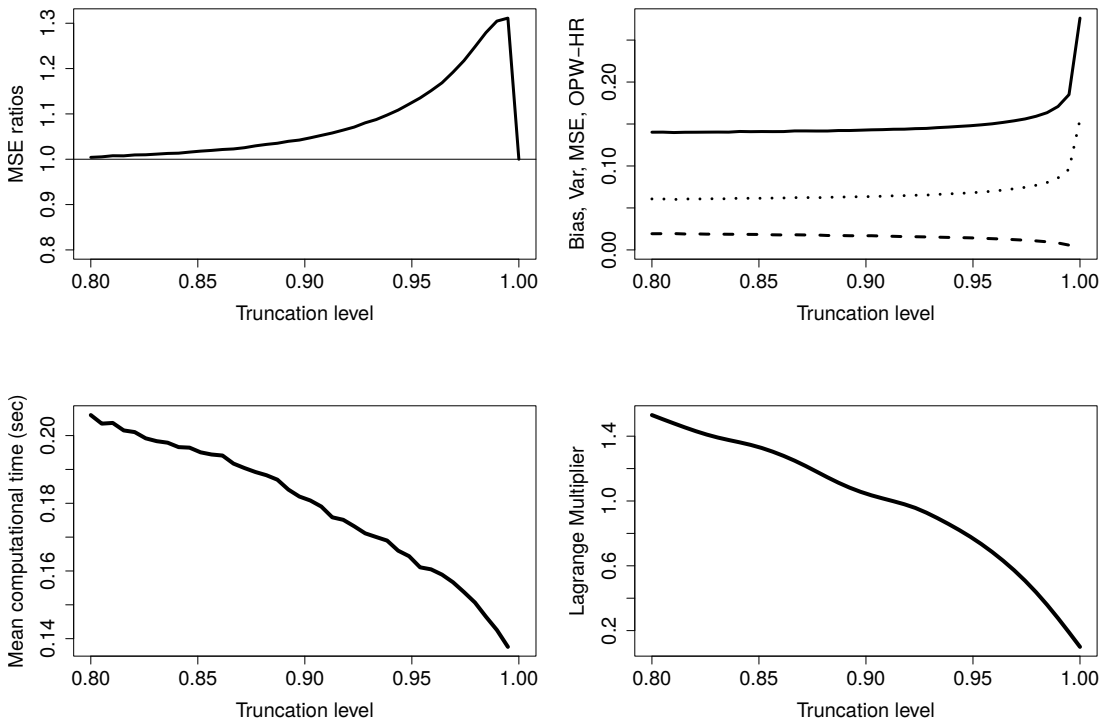

Figure 1: Sample size $n = 200$. Top-left panel: ratio between the observed mean squared error of the estimated hazard ratio that uses truncated weights and that of the estimated hazard ratio that uses OPW across truncation levels. Top-right panel: mean squared error (solid line), variance (dotted), and bias (dashed) of the estimated hazard ratio that uses OPW across truncation levels. The value between the 98th and the 99.5th percentile for which the MSE ratio is largest is the 99.5th percentile. Bottom-left: mean computational time in seconds to solve the quadratic problem across levels of truncation. Bottom-right panel: mean standardized Lagrange multiplier associated with constraint (3.2) across truncation levels.

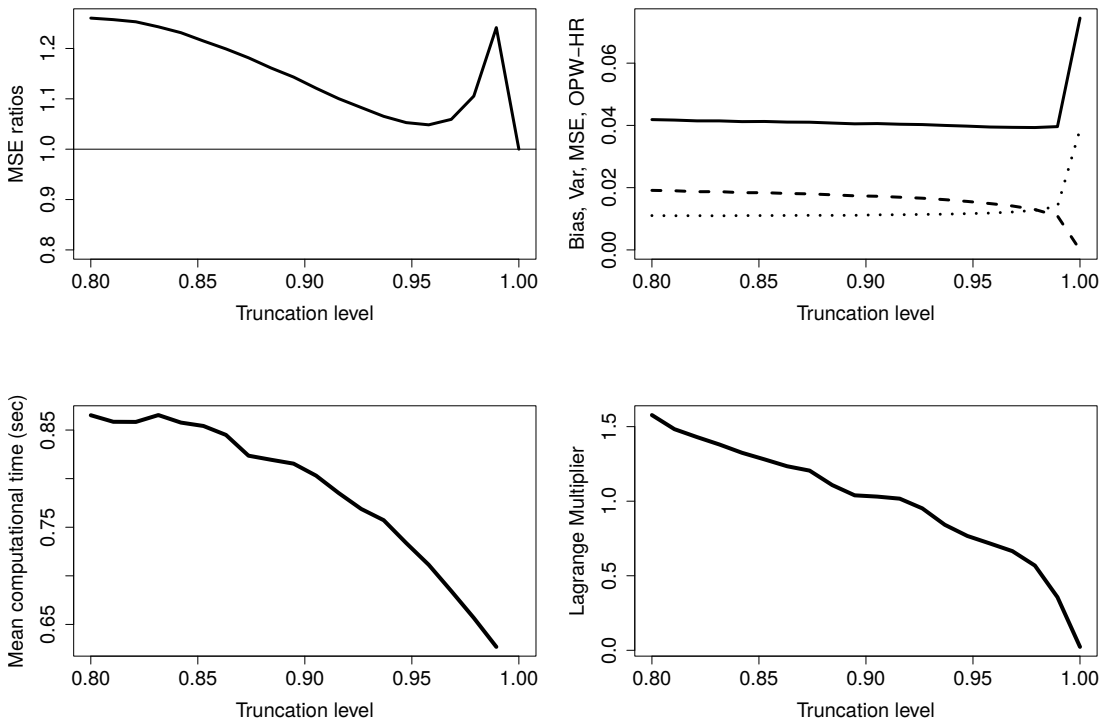

Figure 2: Sample size $n = 1,000$. Top-left panel: ratio between the observed mean squared error of the estimated hazard ratio that uses truncated weights and that of the estimated hazard ratio that uses OPW across truncation levels. Top-right panel: mean squared error (solid line), variance (dotted), and bias (dashed) of the estimated hazard ratio that uses OPW across truncation levels. The value between the 98th and the 99.5th percentile for which the MSE ratio is largest is the 99th percentile. Bottom-left: mean computational time in seconds to solve the quadratic problem across levels of truncation. Bottom-right panel: mean standardized Lagrange multiplier associated with constraint (3.2) across truncation levels.

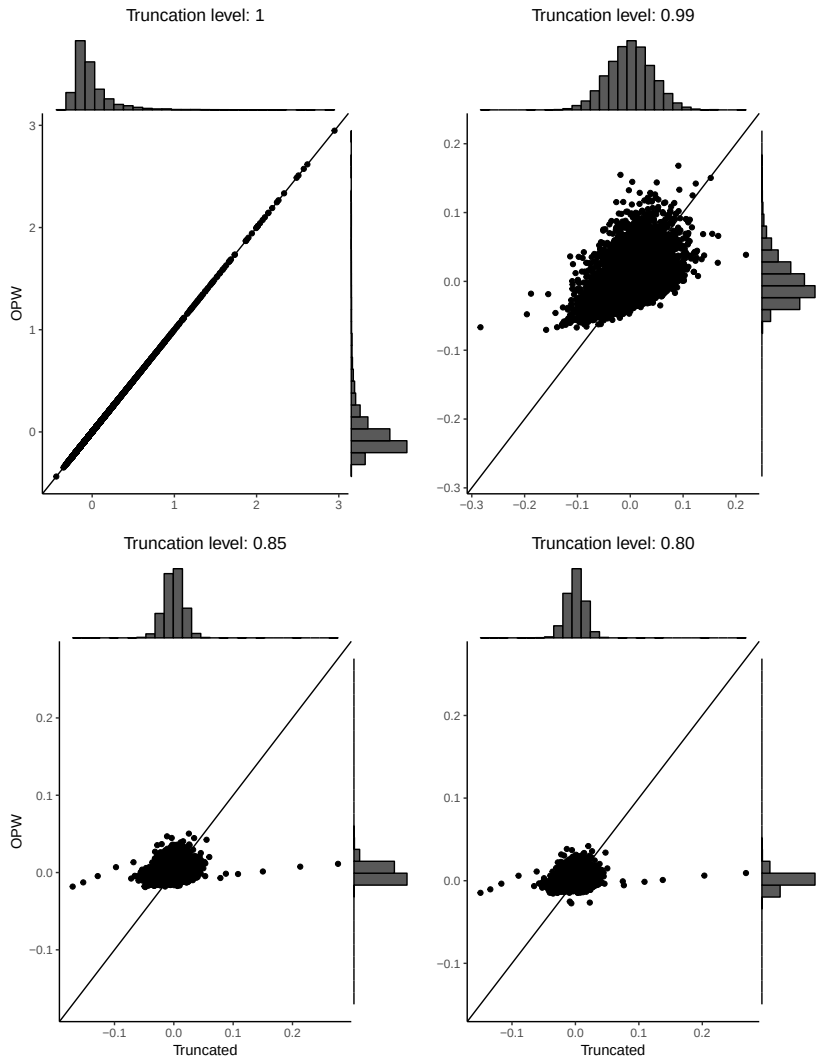

Figure 3: Scatter-plots and histograms of the mean truncated weights, (X-axis), and the mean OPW, (Y-axis), across simulations when $n = 1,000$ for each of the four thresholds, 1, 0.99, 0.85 and 0.80. Weights were first scaled to have mean 0 and variance 1 and then log-transformed.

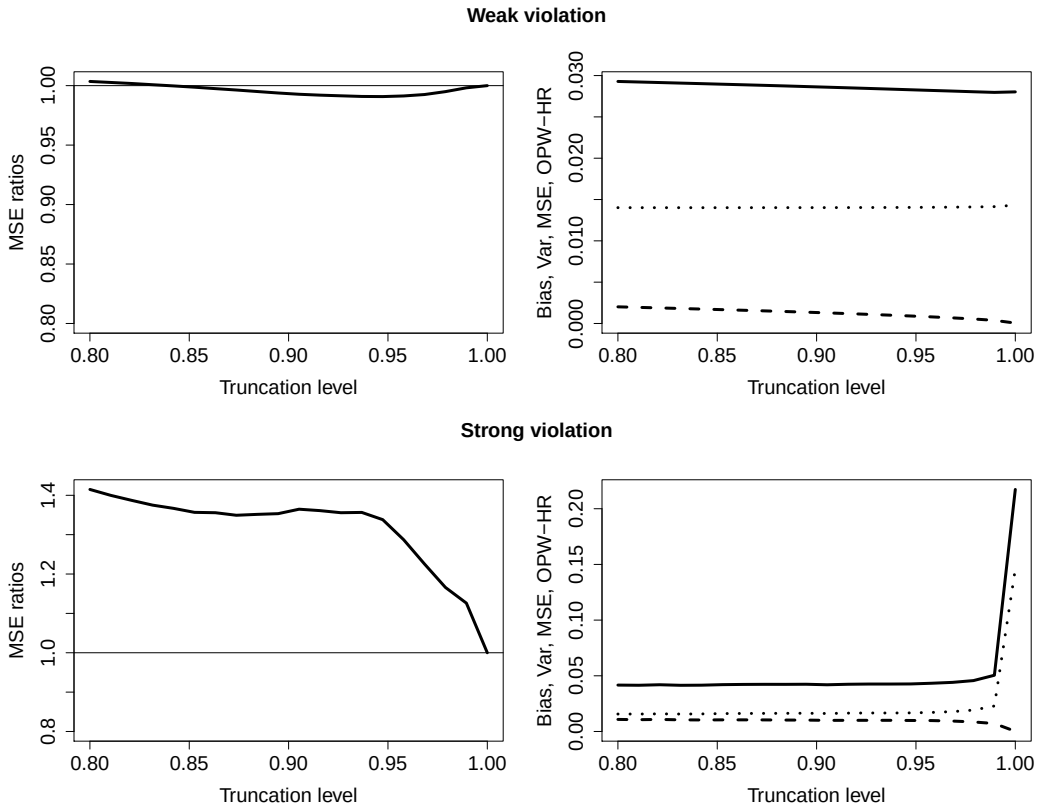

Figure 4: Left panels: MSE ratios between the hazard ratio estimated with truncated weights and that estimated with OPW across truncation levels when $n = 1,000$ under the scenarios of weak and strong violations of the positivity assumption. Right panels: mean squared error (solid line), variance (dotted), and bias (dashed) of the estimated hazard ratio that uses OPW across truncation levels under the scenarios of weak and strong violations of the positivity assumption.

5 HIV treatment initiation on time to death

The HIV epidemic is a leading global burden with major economic and social consequences. Drug injection is responsible for more than 10% of all HIV infections globally Mathers et al. (2008). Consequently, the efficacy of the HIV treatment is of primary concern when treating people who inject drugs (PWID). Several studies have shown the beneficial effect of HIV treatment among PWID Wood et al. (2008); Mathers et al. (2010). We evaluated the effect of HIV treatment initiation on time to death among PWID. To control for time-dependent confounding and informative censoring we used OPW obtained by solving (3.1)-(3.3). We computed the set of target weights as the product between the inverse probability of treatment and censoring weights Robins et al. (2000). As discussed in Section 3.1, we truncated the set of target weights at different truncation levels, computed the variance of the resulting truncated weights and used it as a value for ξ in constraint (3.2).

5.1 Study population

We used prospective observational data from the Swedish InfCare HIV registry Sönnerborg (2017), which contains socio-demographical, clinical and virological information, collected longitudinally from all clinics that treat people living with HIV. The number of people diagnosed between 1987 and 2017 in Sweden was 10,015. Our study was restricted to those who were alive, HIV treatment-naive and under follow-up after January 1996, when HIV treatment became readily available in the country. We excluded 1,055 people who had both their first and last visit before January 1996 (due to emigration or death) and 1,187 who started HIV treatment before January 1996. The baseline visit was set equal to the first available visit for each person. For those enrolled in the HIV monitoring program before January 1996, it was set at the first available visit after January 1996. People living with HIV were

monitored and visited repeatedly from baseline onward contributing from a minimum of 2 to a maximum of 102 visits. At each visit, data on socio-demographical characteristics, type of HIV treatment, laboratory measurements including absolute CD4 cell count and HIV-RNA load were collected. HIV treatment was defined as a combination of at least 3 drugs, classified in 4 major categories: based on non-nucleoside reverse-transcriptase-inhibitors, ritonavir-boosted protease inhibitors, protease inhibitors, and others. Out of the 7,773 people, 459 lacked information on absolute CD4 cell count, 199 had only one absolute CD4 cell count observation, and 1,110 did not have sufficient information on the route of infection. We considered only people living with HIV infected by injecting drugs. The final sample was comprised of 538 treatment-naive PWID and a total of 9,247 clinical visits.

5.2 Treatment and censoring models

We used logistic regression to estimate the set of target weights that control for time-dependent confounding and informative censoring. We used time-invariant and time-dependent confounders to construct the set of stable inverse probability of treatment weights as shown in (2.2). Specifically, we identified the following variables as potential time-invariant confounders of the effect of HIV treatment initiation on time to death: baseline absolute CD4 cell count ($<$200, 200-350, 350-500, and $>$500 cells/mL); baseline HIV-RNA viral load ($\leq$100.000 vs $>$100.000 copies/mL) ; age at baseline (0-30, 31-40, 41-50, and $>$50 years); gender (female vs male); country of birth (Sweden vs. outside Sweden); type of HIV treatment regimen (4 drug categories) and calendar year of HIV treatment initiation. We considered the following potential time-dependent confounders: absolute CD4 cell count, modelled as cubic splines with 3 knots placed at the 25th, 50th and 75th percentiles, cumulative follow-up time, modelled as a cubic splines with 5 knots

at 5th, 25th, 50th, 75th and 95th percentiles, undetectable HIV-RNA viral load and HIV treatment at previous time points. Undetectable HIV-RNA viral load was considered undetectable if it was lower than 50 copies/mL. We constructed the set of inverse probability of censoring weights similarly. Finally, we obtained the set of target stabilized weights as the product between inverse probability of treatment and censoring weights.

5.3 Results

We considered the following MSCM to evaluate the effect of HIV treatment on time to death among PWID,

$$\lambda_{i,k}(t|\overline{A}_i^{(k)}, V_i) = \lambda_0(t)\exp(\beta_1 A_i^{(k)} + \beta_2 A_i^{(k-1)} + \beta_3 V_i) \tag{5.1}$$

where V_i was the baseline absolute CD4 cell count for each PWID. We estimated the MSCM in (5.1) by a weighted Cox proportional hazard model. The unweighted estimated hazard ratio was equal to HR= 1.65 with a robust estimate for the standard error equal to 0.36, suggesting the presence of confounding. When using the set of target weights constructed as previously described, the estimated hazard ratio was equal to 0.68, suggesting a protective effect of the HIV treatment on time to death. The standard error was equal to 0.74, more than twice that of the unweighted analysis. In particular, when analyzing the distribution of the target weights, few subjects (n=2) were assigned a weight of more than 500, showing a possible practical violation of the positivity assumption, although not large. To alleviate the presence of extreme weights we computed the set of OPW and used it to estimate the hazard ratio. Specifically, we considered a grid of truncation levels between 0.8 and 1 and computed the truncated weights. We obtained the set of OPW by setting ξ equal to the variance of the truncated weights for each of the considered trun-

cation levels. When the truncation level was equal to the 99.5th percentile of the target weight distribution, the set of OPW had a minimum value of 0.86 and a maximum value of 27. Figure 5 shows the value of the estimated hazard ratio for the risk of death among PWID and 95% confidence interval across the considered truncation levels. Similarly to our simulations results, we based the conclusions of this study on an estimated HR of 0.71 (95% CI 0.19-2.73, standard error equal to 0.68), obtained by using OPW with ξ set to be equal to the variance of the truncated weights at 99.5th level. We concluded that with adequate support, PWID can benefit from HIV treatment.

6 Conclusions

In this paper, we introduced OPW to the estimation of causal effects of time-varying treatments on survival outcomes with MSCM under practical violation of the positivity assumption. Xiao et al. Xiao et al. (2013) and Cole and Hernán Cole and Hernán (2008) suggested truncating the weights at high percentiles of their observed sample distribution. In our simulations, OPW outperformed the truncated weights across all the considered truncation levels, especially at high percentiles. The results were similar in both small and large samples. In addition, the results showed that OPW were generally more narrowly distributed than truncated weights across all considered threshold levels and that OPW outperformed truncated weights across all scenarios of practical violation of the positivity assumption, especially under strong violation. This suggests that OPW may be used instead of truncated weights regardless of the sample size and the strength of practical positivity violation. By using OPW, we showed the beneficial effect of treatment on time to death among people in Sweden who live with HIV after being infected by injecting drugs.

We considered MSCM, but other methods, such as pooled logistic regression, can also

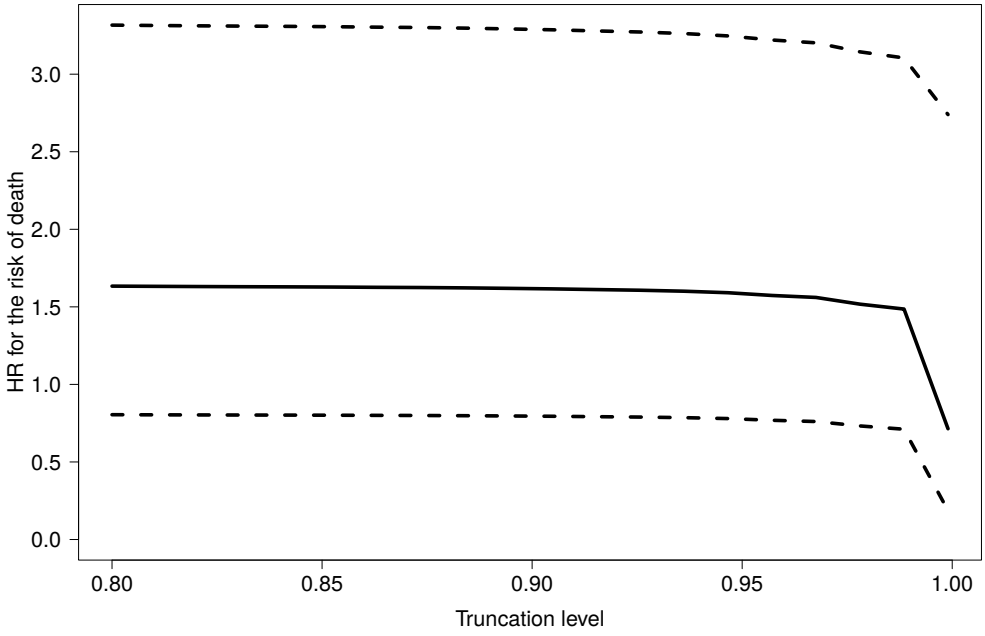

Figure 5: Estimated hazard ratio for the risk of death and 95% confidence interval comparing treated vs. untreated individuals across levels of truncation of the target stabilized weights.

be used Robins et al. (2000). Different methods to estimate the standard error may also be applied, such as the bootstrap. In any given setting these may be preferable to the robust estimator we used in the present paper. The optimization problem (3.1)-(3.3) and its interpretation remain unchanged whichever estimator is used. We show this by providing additional simulations as Supporting Information.

We derived the target weights by using logistic regression. However, a number of alternative techniques have been proposed Karim et al. (2017); Lee et al. (2010, 2011). We considered scenarios where the treatment and censoring models were well specified. When they are suspected to be misspecified, Karim et al. Karim et al. (2017) suggested using boosted regression and classification trees. These can be used to estimate the set of target weights employed in (3.1)-(3.3).

The convex optimization problem (3.1)-(3.3) can be solved by using existing software, like `gurobi`, `quadprog`, `Ipoptr`, and `nloptr` packages in `R`. The sample size has an impact on the computation time of the proposed method. For instance, in our simulations the average time was 0.2 seconds with $n = 200$ and about 1 second with $n = 1,000$. Decreasing ξ increases the computational time and may impact the feasibility of the problem. With small values of ξ, an optimal solution may not exist. In this case, we suggest increasing the value of ξ.

Future work may focus on extensions and applications of OPW to a variety of other settings. For example, they may prove useful when comparing dynamic treatment regimes, where treatment decisions are made based on the time-varying state of individual patients and weights are applied to control for time-dependent confounding, and informative and artificial censoring Hernán et al. (2006, 2009). Further work may improve the robustness to misspecification of the treatment model and violations of the positivity assumption.

References

Athey, S., G. W. Imbens, and S. Wager (2016). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*.

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

Blackwell, M. (2013). A framework for dynamic causal inference in political science. *American Journal of Political Science 57*(2), 504–520.

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

Cole, S. R., M. A. Hernán, J. B. Margolick, M. H. Cohen, and J. M. Robins (2005). Marginal structural models for estimating the effect of highly active antiretroviral therapy initiation on CD4 cell count. *American Journal of Epidemiology 162*(5), 471–478.

Daniel, R., S. Cousens, B. De Stavola, M. Kenward, and J. Sterne (2013). Methods for dealing with time-dependent confounding. *Statistics in Medicine 32*(9), 1584–1618.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis 20*(1), 25–46.

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

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

Hernán, M. A., E. Lanoy, D. Costagliola, and J. M. Robins (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & Clinical Pharmacology & Toxicology 98*(3), 237–242.

Hernán, M. A., M. McAdams, N. McGrath, E. Lanoy, and D. Costagliola (2009). Observation plans in longitudinal studies with time-varying treatments. *Statistical Methods in Medical Research 18*(1), 27–52.

Hernan, M. A. and J. M. Robins (2010). *Causal inference*. CRC Boca Raton, FL.

HIV-Causal Collaboration (2011). When to initiate combined antiretroviral therapy to reduce mortality and AIDS-defining illness in HIV-infected persons in developed countries: an observational study. *Annals of Internal Medicine 154*(8), 509.

HSL (2017). A collection of Fortran codes for large scale scientific computation. `http://www.hsl.rl.ac.uk`.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kang, J. D. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Karim, M. E., J. Petkau, P. Gustafson, H. Tremlett, and T. B. S. Group (2017). On the application of statistical learning approaches to construct inverse probability weights in

marginal structural Cox models: Hedging against weight-model misspecification. *Communications in Statistics-Simulation and Computation 0*(0), 1–30.

Lee, B. K., J. Lessler, and E. A. Stuart (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine 29*(3), 337–346.

Lee, B. K., J. Lessler, and E. A. Stuart (2011). Weight trimming and propensity score weighting. *PloS one 6*(3), e18174.

Mathers, B. M., L. Degenhardt, H. Ali, L. Wiessing, M. Hickman, R. P. Mattick, B. Myers, A. Ambekar, S. A. Strathdee, et al. (2010). HIV prevention, treatment, and care services for people who inject drugs: a systematic review of global, regional, and national coverage. *The Lancet 375*(9719), 1014–1028.

Mathers, B. M., L. Degenhardt, B. Phillips, L. Wiessing, M. Hickman, S. A. Strathdee, A. Wodak, S. Panda, M. Tyndall, A. Toufik, et al. (2008). Global epidemiology of injecting drug use and HIV among people who inject drugs: a systematic review. *The Lancet 372*(9651), 1733–1745.

Neugebauer, R., M. J. van der Laan, M. M. Joffe, and I. B. Tager (2007). Causal inference in longitudinal studies with history-restricted marginal structural models. *Electronic Journal of Statistics 1*, 119.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science 22*(4), 544–559.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools

for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pp. 95–133. Springer.

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

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association 90*(429), 106–121.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Santacatterina, M. and M. Bottai (2017). Optimal probability weights for inference with constrained precision. *Journal of the American Statistical Association 0*(ja), 0–0.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120.

Sönnerborg, A. (2017). InfCare HIV database. `http://infcare.se/hiv/sv/`. Accessed: 2017-08-16.

Stürmer, T., K. J. Rothman, J. Avorn, and R. J. Glynn (2010). Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution - a simulation study. *American Journal of Epidemiology 172*(7), 843–854.

van der Wal, W. M., R. B. Geskus, et al. (2011). Ipw: an R package for inverse probability weighting. *Journal of Statistical Software 43*(13), 1–23.

Wächter, A. and L. T. Biegler (2005, April). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming 106*(1), 25–57.

Wood, E., T. Kerr, M. W. Tyndall, and J. S. Montaner (2008). A review of barriers and facilitators of HIV treatment among injection drug users. *AIDS 22*(11), 1247–1256.

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

Xiao, Y., E. E. Moodie, and M. Abrahamowicz (2013). Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods 2*(1), 1–20.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association 110*(511), 910–922.