Comparison of Weighting Methods in the Estimation of Time-to-Event Causal Effects

Can Meng and Waveley Qiu

BIS 537 Final Project

Fall 2023

1. Introduction

Time-to-event outcomes are of great interest to investigators across a variety of clinical settings, however present analytical difficulties even within the constrains of a randomized controlled trial. In particular, censoring, which is the non-observation of an event yet loss of a subject's data, presents another missing data issue that the investigator has need to consider. This issue is compounded in the setting of a non-randomized or observational study, where exchangeability between treatment groups cannot be assumed of the collected data, and the researcher is left with a complex dual-level missing data issue to reckon with.

Several procedures addressing these two missing data issues in the estimation of time-to-event causal effects have been explored previously. Some of these procedures extend the idea of inverse probability weighting to the setting of survival analysis in constructing a combined treatment and censoring weights (Hernan 2000, Cheng 2022) to approximate exchangeability and estimating the causal estimand through a weighted estimator. Other methods approach the problem by obtaining survival outcomes for each subject through a jackknife procedure (Andersen 2010) and estimate the causal estimand based on those pseudo-observed endpoints. The pseudo-observaions approach has been further improved upon with the employment of propensity score weighting to close the distance between the collected sample and the target population (Li 2021) as well as with the recent proposal of a doubly-robust estimation technique (Wang 2023) to allow for some flexibility with regards to model misspecification.

In this study, we are interested in revisiting the weighting methods and are proposing a new procedure in which model misspecification may be implicitly adjusted for. In doing so, we will first employ the two-arm weighting method established by Cheng et.al (2022), in which weights based on the probability of treatment and the probability of censoring are calculated and used together in estimating the causal estimand of interest. We will then compare its performance to the results obtained from a single-arm sequential weighting method that we have developed, where weights are constructed iteratively, and the final compositely-weighted dataset will then be used to estimate the causal estimand.

The remainder of this paper is organized as follows: in section 2, we introduce the methods developed and utilized in this study; in section 3, we discuss our data generation procedure and simulation design; in section 4, we illustrate our results from the conducted simulation study; and section 5 concludes with a discussion.

2. Methods

2.1 Causal Survival Estimand

While there are several causal estimands that could be of interest, we will be focusing our study on the restricted average causal effect (RACE) for ease of interpretation. In Mao et al. (2018), this was defined as follows:

$$\begin{aligned}\Delta_{RACE} &= \frac{E[\omega(e_i)\min(T_{1i}, t^*)]}{E[\omega(e_i)]} - \frac{E[\omega(e_i)\min(T_{0i}, t^*)]}{E[\omega(e_i)]} \\ &= \int_0^{t^*} S_1(t)dt - \int_0^{t^*} S_0(t)dt\end{aligned}$$

This estimand is interpreted as the average difference in survival time between the treatment and control groups, if both potential outcomes are observed, under the upper bound time restriction of t^*.

In this study, we are interested in the upper bound time restriction of 5 – thus, our estimand will take the following form:

$$\begin{aligned}\Delta_{RACE} &= \frac{E[\omega(e_i)\min(T_{1i}, 5)]}{E[\omega(e_i)]} - \frac{E[\omega(e_i)\min(T_{0i}, 5)]}{E[\omega(e_i)]} \\ &= \int_0^{5} S_1(t)dt - \int_0^{5} S_0(t)dt\end{aligned}$$

2.2 Causal Estimators

In the separate weighting procedure, to incorporate both the inverse probability treatment weights (IPTW) and the inverse probability of censoring weights (IPCW), we employed the usage of the non-parametric weighted Kaplan-Meier estimator defined by Cheng et al. (2022). This estimator is defined as follows:

$$\begin{aligned}\hat{\Delta}(t) &= \hat{S}_1(t) - \hat{S}_0(t) \\ &= \left(1 - \frac{\sum_{i=1}^{n} \omega_{IPTW_i} Z_i \delta_i I(U_i \leq t) \omega_{IPCW_i}}{\sum_{i=1}^{n} \omega_{IPTW_i} Z_i}\right) - \left(1 - \frac{\sum_{i=1}^{n} \omega_{IPTW_i} (1 - Z_i) \delta_i I(U_i \leq t) \omega_{IPCW_i}}{\sum_{i=1}^{n} \omega_{IPTW_i} (1 - Z_i)}\right)\end{aligned}$$

where T_i^S is the event time of interest, T_i^C is the censoring time, $U_i = min(T_i^S, T_i^C)$ is the observed time in study, δ_i is the censoring indicator, Z_i is the treatment indicator, $I(U_i \leq t)$ is the indicator for the observed time being less than the upper limit of time t, and ω_{IPTW_i} and ω_{IPCW_i} are IPTW and IPCW respectively. Then, our estimator would be defined as follows, at $t = 5$:

$$\hat{\Delta}(5) = \left(1 - \frac{\sum_{i=1}^{n} \omega_{IPTW_i} Z_i \delta_i I(U_i \leq 5) \omega_{IPCW_i}}{\sum_{i=1}^{n} \omega_{IPTW_i} Z_i}\right) - \left(1 - \frac{\sum_{i=1}^{n} \omega_{IPTW_i} (1 - Z_i) \delta_i I(U_i \leq 5) \omega_{IPCW_i}}{\sum_{i=1}^{n} \omega_{IPTW_i} (1 - Z_i)}\right)$$

The estimators used in the sequentially weighted procedure will similarly be weighted Kaplan-Meier estimators – however, these will be distinct from the previously defined estimator in that there will be the need to adjust for one set of weights only. The estimator will thus have the following form, both generally and at $t = 5$:

$$\begin{aligned}\hat{\Delta}(t) &= \hat{S}_1(t) - \hat{S}_0(t) \\ &= \left(1 - \frac{\sum_{i=1}^{n} \omega_{IPTW \circ IPCW_i} Z_i \delta_i I(U_i \leq t)}{\sum_{i=1}^{n} \omega_{IPTW \circ IPCW_i} Z_i}\right) - \left(1 - \frac{\sum_{i=1}^{n} \omega_{IPTW \circ IPCW_i} (1 - Z_i) \delta_i I(U_i \leq t)}{\sum_{i=1}^{n} \omega_{IPTW \circ IPCW_i} (1 - Z_i)}\right) \\ \hat{\Delta}(5) &= \hat{S}_1(5) - \hat{S}_0(5) \\ &= \left(1 - \frac{\sum_{i=1}^{n} \omega_{IPTW \circ IPCW_i} Z_i \delta_i I(U_i \leq 5)}{\sum_{i=1}^{n} \omega_{IPTW \circ IPCW_i} Z_i}\right) - \left(1 - \frac{\sum_{i=1}^{n} \omega_{IPTW \circ IPCW_i} (1 - Z_i) \delta_i I(U_i \leq 5)}{\sum_{i=1}^{n} \omega_{IPTW \circ IPCW_i} (1 - Z_i)}\right)\end{aligned}$$

where $\omega_{IPTW \circ IPCW_i}$ is the composite treatment and censoring weight (and likewise for the composite censoring and treatment weight) and all other notation is the same as defined previously.

2.3 Weighting Procedures

Critical to the construction of the causal estimators are the weighting procedures underlying the construction of the inverse probability weights of both treatment assignment and censoring indication. The goal for both of the procedures that are the subject of investigation in this study is covariate balance between treatment and control groups, such that the study sample would mimic the distribution of the theoretical larger population that it is drawn from.

All weights constructed in this study were inverse probability weights – in particular, for the construction of a weight on indicator variable R, the weights w_{IPRW_i} for a given subject i were constructed as follows:

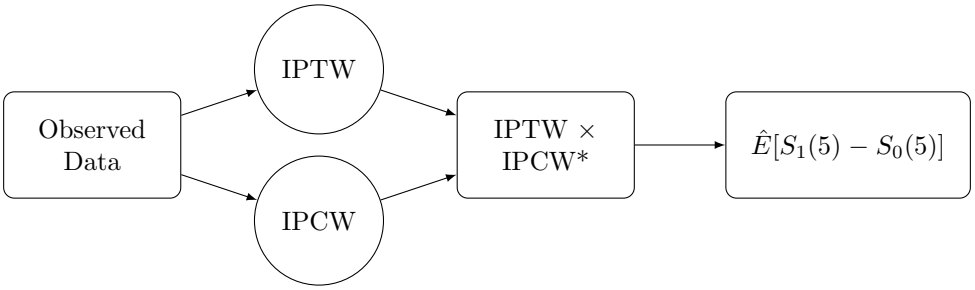
$$\mathbb{E}[R|\mathbf{X}] = e_R(\mathbf{X}), \quad \text{where } \mathbf{X} \text{ is the vector of covariate values}$$

$$w_{IPRW_i} = \begin{cases} \frac{1}{e_R(\mathbf{X})}, & \text{if R} = 1 \\ \frac{1}{1-e_R(\mathbf{X})}, & \text{if R} = 0 \end{cases}$$

In the estimated propensity score models, all available baseline covariates (which does exclude treatment assignment and censoring indication) were included as model predictors, irrespective of how the true propensity score model was generated. Though this by default produces a misspecified model, our intent was to replicate and examine situations that may occur in practice, where the data generation procedure is unknown and it may seem reasonable to the investigator to include all collected baseline characteristics. We will explore the impact of this decision in the Discussion section of this paper.

2.3.1 Separate Weighting

The first weighting procedure is to estimate IPTW and IPCW separately. Then the two weights are used together in the estimator described in the previous section to estimate our target estimand (see the diagram below).

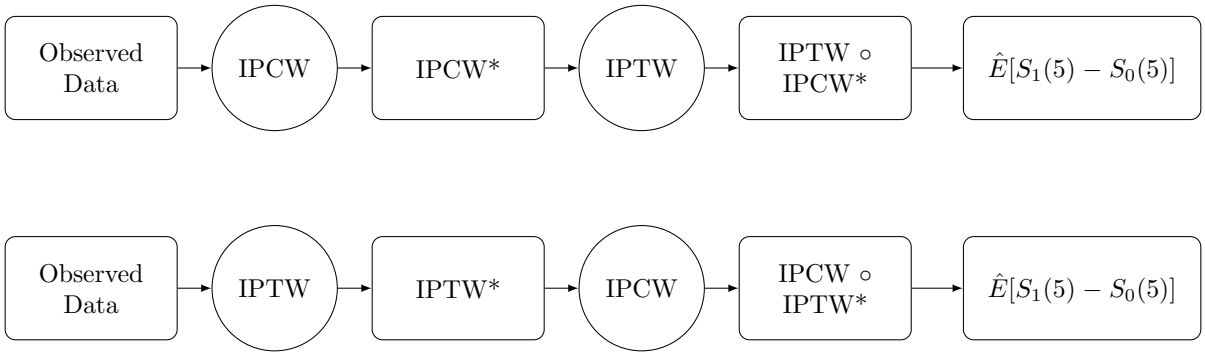


In the separate weighting procedure, IPTW was estimated solely using logistic regression, and IPCW was estimated using both logistic regression and Cox-Exponential regression. The logistic regressions were fitted using `glm` function and the Cox-Exponential models were fit using `ipcw` from the `riskRegression` R package.

2.3.2 Sequential Weighting

In this second weighting procedure, weights are first constructed based on either treatment assignment or censoring indication propensity scores. After this, the weighted dataset would be used to generate weights based on the indication that was not used in the first step of weighting. The causal estimand would be estimated using the second set of weights applied to to the weighted dataset from the first step of weighting.

The weighting processes were conducted as depicted in the following diagram, where circles indicate where weighting occurred and boxes marked with asterisks indicate pseudo-populations:



We have considered the utility of both indicator variables (treatment assignment and censoring indication) as weighting starting points because the relationship between censoring indication and treatment assignment causes it to be difficult to favor one sequential weighting pathway over the other. Some may prefer the sequential weighting procedure that calculates treatment weights first, on the basis of observational temporality – however, while it is true that treatment assignment is observed by the researcher prior to censoring time, if the subject indeed is censored, it seems that the timeline of the conduct of the actual study may not necessarily be the best guideline in determining the order of how weights are calculated, especially if it is assumed that censoring time and treatment assignment are independent of each other, as is commonly assumed and has been maintained in this study. In such a case, when censoring time is a quantity that is generated without respect to treatment assignment, the censoring indicator variable

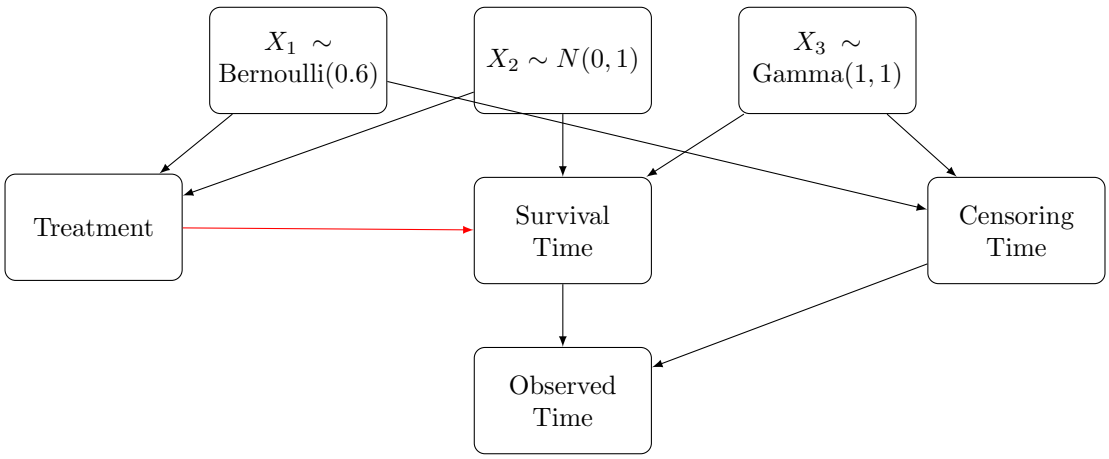
would be related to treatment assignment indirectly, through the effect that the assigned treatment would have on survival time; this is a rather complex relationship to evaluate

Thus, in this study, we have considered and presented results for both options of sequential weighting, allowing room for investigators interested in this method of inverse probability weighting adjustment to select the method most suitable for the parameters of their studies.

In the implementation of both procedures, the treatment weights were obtained through a logistic regression in `R`, using the `glm` function in base `R` with a logit link function, and the censoring weights were obtained through two Cox-Exponential regressions, which were separated by treatment group, through the `ipcw` function in the `riskRegression` package. In the establishment of the IPTW $\circ$ IPCW procedure, the censoring weights were directly identified in the `glm` function call. In the establishment of the IPCW $\circ$ IPTW procedure, as the `ipcw` function does not take in weights, the treatment-weighted dataset was expanded, using the `expandRows` function from the `splitstackshape` package, according to the calculated treatment weights in order to obtain the second set of weights for censoring indication.

3. Data Generation

We will generate data in accordance with the following relationships, and the causal effect of interest is highlighted in red:



Each component will be described in further detail in the following sections.

3.1 Covariates

As described in the previous diagram, we will be utilizing three covariates in this simulation study: X_1, X_2, X_3. These covariates have the following distributions: $X_1 \sim$ Bernoulli(0.6), $X_2 \sim N(0, 1)$, $X_3 \sim$ Gamma$(1, 1)$ and will explicitly be defined to not be time dependent. These distributions were selected to represent different kinds of covariates that may appear in a real-world study and also to a variety of distributional behaviors.

In addition, these three covariates will all affect the critical measures of interest in this study (i.e., treatment, survival time, and censoring time). However, any given measure will only be affected by two of the three covariates and no two measures will be affected by the same combination of covariates.

3.2 True Propensity Score Model

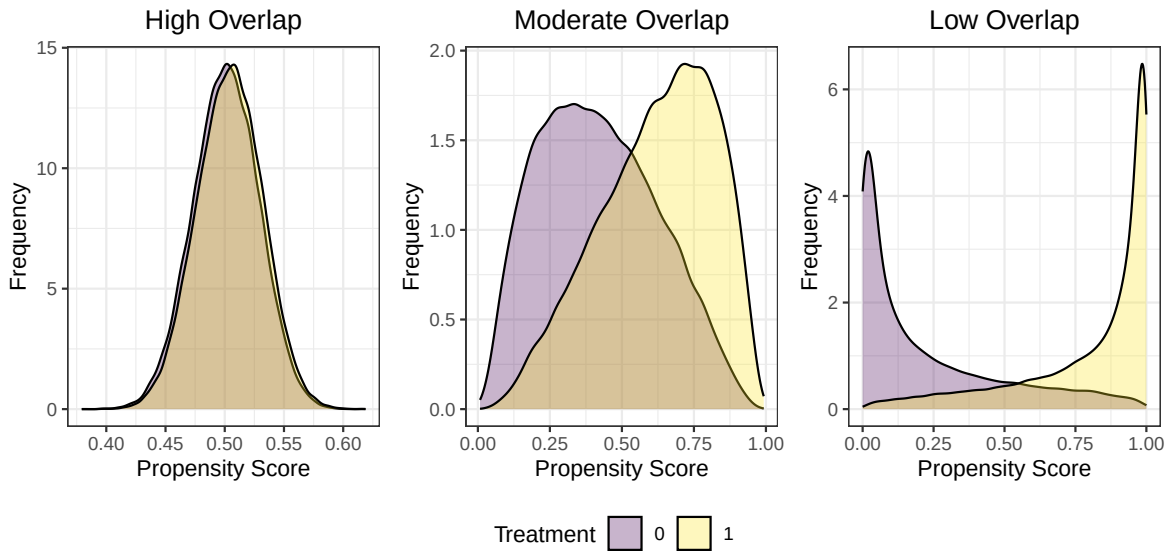
The true propensity score model will be defined using the following logistic regression, taking in covariates X_1 and X_2:

$$g^{-1}(E[Z = 1|\mathbf{X}_i]) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i},$$

$$\text{where } g(\mathbf{X_i}) = \frac{\exp(\boldsymbol{\beta}\mathbf{X}_i^T)}{1 + \exp(\boldsymbol{\beta}\mathbf{X}_i^T)} = E[Z = 1|\mathbf{X}_i] = e(\mathbf{X}_i)$$

In our simulation, β_1, β_2 was held fixed at $\beta_1 = \beta_2 = 0.1$ for the setting of high overlap of overlap, $\beta_1 = \beta_2 = 1$ for the setting of medium overlap, and $\beta_1 = \beta_2 = 3$ for the setting of low overlap. For each of these settings, the β_0 that would produce an expected treatment proportion of 55% was be obtained through a numerical search over an interval.

The distribution of propensity scores according to each of these settings can be visualized with the following plots, demonstrating that the levels of overlap between treatment groups have indeed been achieved:



3.3 True Survival Time Model

The true survival time model will be defined using the following Cox-Weibull mode, taking in covariates X_1 and X_3:

$$
\begin{aligned}
h(t|\mathbf{X}_i) &= h_0(t)\exp(L_i), \\
\text{where } h_0(t) = \lambda\nu t^{\nu-1}&, \text{ and } L_i = a_0 Z_i + a_1 X_{1i} + a_3 X_{3i} \\
S(T_i^S = t) &= \exp(-\lambda t^v \exp(L_i)) \\
&= 1 - F(t)
\end{aligned}
$$

The survival time for subject i will then be drawn from:

$$
\begin{aligned}
T_i^S &= \left(\frac{-\log(u_i^S)}{\lambda\exp(L_i)}\right)^{1/\nu}, \\
&\text{where } u_i^S \sim \text{Unif}(0,1)
\end{aligned}
$$

From this model, the treatment effect will be determined by α_0. Two settings of α_0 were selected to correspond with a low and high treatment effect respectively.

To obtain the true treatment effects from each of these settings, we started with the conditional distribu-

tions of survival time that would be obtained from the defined Cox-Weibull model:

$$\begin{aligned} T(1)\Big|(X_1, X_3) &\sim h_{Z=1}(t|X_1, X_3) = \lambda v t^{v-1} \exp(\alpha_0 + \alpha_1 X_1 + \alpha_3 X_3) \\ &= \lambda^* \nu t^{\nu-1}, \text{ where } \lambda^* = \lambda \exp(\alpha_0 + \alpha_1 X_1 + \alpha_3 X_3) \\ T(0)\Big|(X_1, X_3) &\sim h_{Z=0}(t|X_1, X_3) = \lambda v t^{v-1} \exp(\alpha_1 X_1 + \alpha_3 X_3) \\ &= \lambda^* \nu t^{\nu-1}, \text{ where } \lambda^* = \lambda \exp(\alpha_1 X_1 + \alpha_3 X_3) \end{aligned}$$

Our target estimand is the restricted average causal effect at time = 5, or $E[S_1(5) - S_0(5)]$, which is unconditional on (X_1, X_3). By the law of total expectations and iterated expectations, we see the following:

$$\begin{aligned} E[S_{T_i}(5)] &= E\Big[E\Big[I(T_i \geq 5)|X_1, X_3\Big]\Big] \\ &= \int_{X_1, X_3} P\Big[T_i \geq 5|X_1 = x_1, X_3 = x_3]P(X_1 = x_1, X_3 = x_3)d\mu(x_1, x_3) \\ &= \int_{X_1, X_3} P\Big[T_i \geq 5|X_1 = x_1, X_3 = x_3]P(X_1 = x_1)P(X_3 = x_3)d\mu(x_1, x_3) \end{aligned}$$

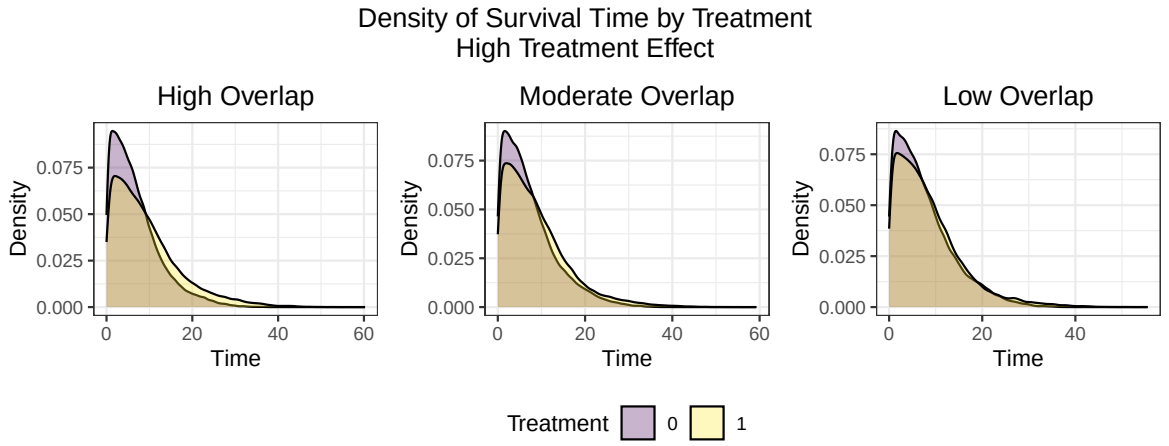
Let $X_{11}, ..., X_{1m} \sim P(X_1)$ and $X_{31}, ..., X_{3m} \sim P(X_3)$. Then, for a sufficiently large m, we note:

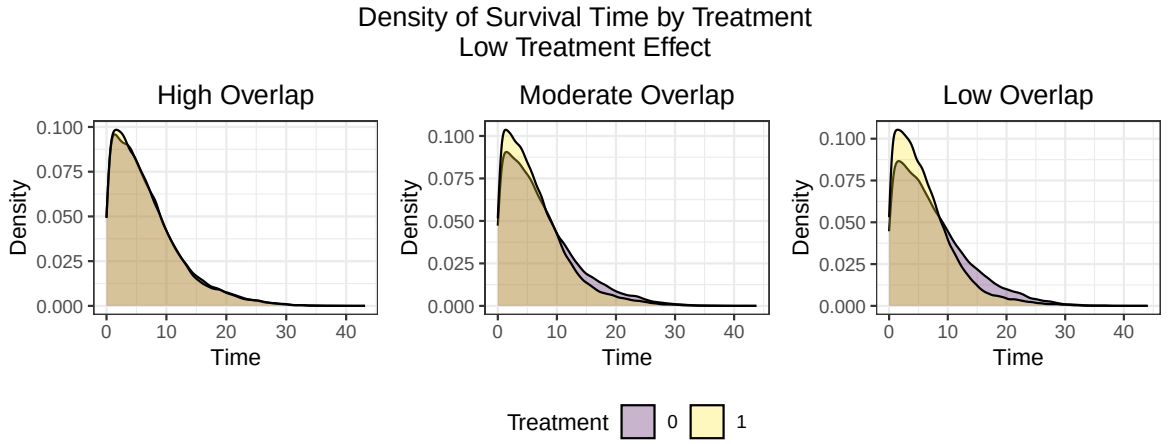
$$\begin{aligned} P[T_i \geq 5] &= \int_{X_1, X_3} P\Big[T_i \geq 5|X_1 = x_1, X_3 = x_3]P(X_1 = x_1)P(X_3 = x_3)d\mu(x_1, x_3) \\ &\approx \sum_{j=1}^{m} P[T_i \geq 5|X_1 = x_{1j}, X_3 = x_{3j}] \end{aligned}$$

The true value estimates of $S_1(5) = P[T(1) \geq 5]$ and $S_0(5) = P[T(0) \geq 5]$ will be calculated by taking two samples of size $m = 1000000$ from the distributions of X_1 and X_3 to simulate the distributional behavior of each of these random variables, and sum the computed conditional probabilities of $P[T(1) \geq 5|(X_{1j}, X_{3j})]$ and $P[T(0) \geq 5|(X_{1k}, X_{3k})]$ for each sample j and k, where $j = k = m$.

The true RACE values that we computed from this procedure were 0.12 for the "high" treatment effect setting and 0.014 for the "low" treatment effect setting.

The distribution of survival time according to each of these settings and by levels of covariate overlap between treatment groups can be visualized with the following plots:





3.4 True Censoring Model

The censoring model will be defined using the following exponential model, taking in covariates X_2 and X_3, and is generated independently of T_i^C, T_i^S, and Z_i:

$$T_i^C \sim \text{Exponential}(\lambda \exp(K_i)),$$
$$\text{where } K_i = \gamma_0 + \gamma_2 X_{2i} + \gamma_3 X_{3i}$$

Algorithmically, the censoring time for subject i will be drawn from:

$$T_i^C = \frac{-\log(u^C)}{\lambda \exp(K_i)},$$
$$\text{where } u^C \sim \text{Unif}(0, 1)$$

For subject i, the observed time, T^{obs} is the minimum of T_i^C and T_i^S, or $T^{obs} = \min(T_i^C, T_i^S)$. In addition,

the censoring indicator for subject i, given the survival time T_i and censoring time T^C, will be assigned as follows:

$$\delta_i(T_i^S, T_i^C) = \begin{cases} 1, & \text{where } T_i^S > T_i^C \\ 0. & \text{where } T_i^S \leq T_i^C \end{cases}$$

It is critical to note that although T_i^C is by definition independent of T_i^S and Z_i, the censoring indication, δ_i is not.

In a similar fashion to how the model intercept was selected in the true propensity score model in order to fix a treatment proportion in expectation, after fixing $\gamma_2 = 2$ and $\gamma_3 = 4$, a numerical search procedure was also run to select a γ_0 so that the proportion of censored subjects could be varied across two levels, corresponding to low (25%) and high (50%) levels of censoring. The selected intercepts varied according to the magnitude of treatment effect, as well as the defined censoring level, consequentially resulting in four defined γ_0 values corresponding to the four treatment effect and censoring level categories. These were as follows: $\gamma_0 = 1.04$ for the high treatment effect with low censoring setting, $\gamma_0 = 1.2$ for the low treatment effect with low censoring setting, $\gamma_0 = 3.35$ for the high treatment effect with high censoring setting, and $\gamma_0 = 3.71$ for the low treatment effect with high censoring setting.

4. Results

In the conduct of this simulation study, data were randomly generated from the distributions described in the previous section to simulate random samples of size 1000. These datasets were then weighted through one of the four described weighting procedures, covariate balance was assessed through absolute standard difference (ASD), and weighted Kaplan-Meier estimates were computed to estimate $\Delta_{RACE}(5)$. Bootstrap replicates were drawn to compute the relative bias and empirical 95% coverage rates, and their results were summarized in tabular as well as graphical formats.

4.1 Separate Weighting

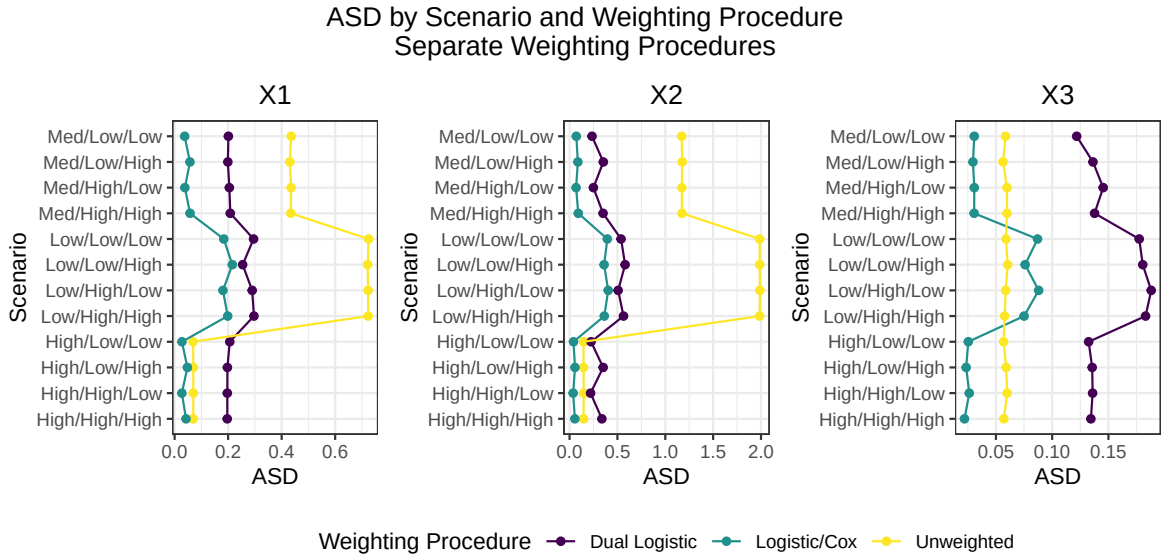
The following table and plot describe the covariate balance results from the two separate weights weighting procedures:

Table 1: Separate Weighting Covariate Balance Results (ASD)[a]

	X1			X2			X3		
Scenario[b]	Unweighted	Dual Logistic	Logistic/Cox	Unweighted	Dual Logistic	Logistic/Cox	Unweighted	Dual Logistic	Logistic/Cox
High/High/Low	0.0698	0.1970	0.0272	0.1477	0.2195	0.0377	0.0600	0.1359	0.0263
Med/High/Low	0.4353	0.2046	0.0385	1.1735	0.2493	0.0682	0.0598	0.1454	0.0308
Low/High/Low	0.7229	0.2899	0.1808	1.9886	0.5076	0.4037	0.0587	0.1880	0.0880
High/Low/Low	0.0691	0.2062	0.0274	0.1467	0.2248	0.0413	0.0568	0.1324	0.0254
Med/Low/Low	0.4358	0.2011	0.0375	1.1714	0.2348	0.0707	0.0586	0.1218	0.0308
Low/Low/Low	0.7250	0.2951	0.1837	1.9858	0.5381	0.3948	0.0589	0.1774	0.0870
High/High/High	0.0698	0.1967	0.0427	0.1464	0.3377	0.0553	0.0570	0.1344	0.0220
Med/High/High	0.4343	0.2082	0.0579	1.1740	0.3501	0.0911	0.0598	0.1377	0.0307
Low/High/High	0.7235	0.2966	0.1985	1.9852	0.5636	0.3629	0.0577	0.1830	0.0751
High/Low/High	0.0702	0.1973	0.0478	0.1462	0.3511	0.0566	0.0591	0.1356	0.0235
Med/Low/High	0.4306	0.1990	0.0574	1.1767	0.3530	0.0869	0.0564	0.1363	0.0296
Low/Low/High	0.7214	0.2543	0.2158	1.9860	0.5801	0.3608	0.0605	0.1805	0.0760

[a] Averaged over 1000 total bootstrapped replicates (100 bootstrap replicates over each of 100 samples)

[b] Scenarios are described by (degree of overlap/treatment effect level/censoring level)



We note that the weighting method using logistic regression for IPTW and Cox model for IPCW outperformed the dual-logistic regression weighting method in terms of balancing baseline covariates, as the logistic/Cox method was consistently associated with lower ASDs than the dual-logistic regression method. Both of these methods produced better covariate balance than what existed in the unweighted data on all covariates excluding X_3.

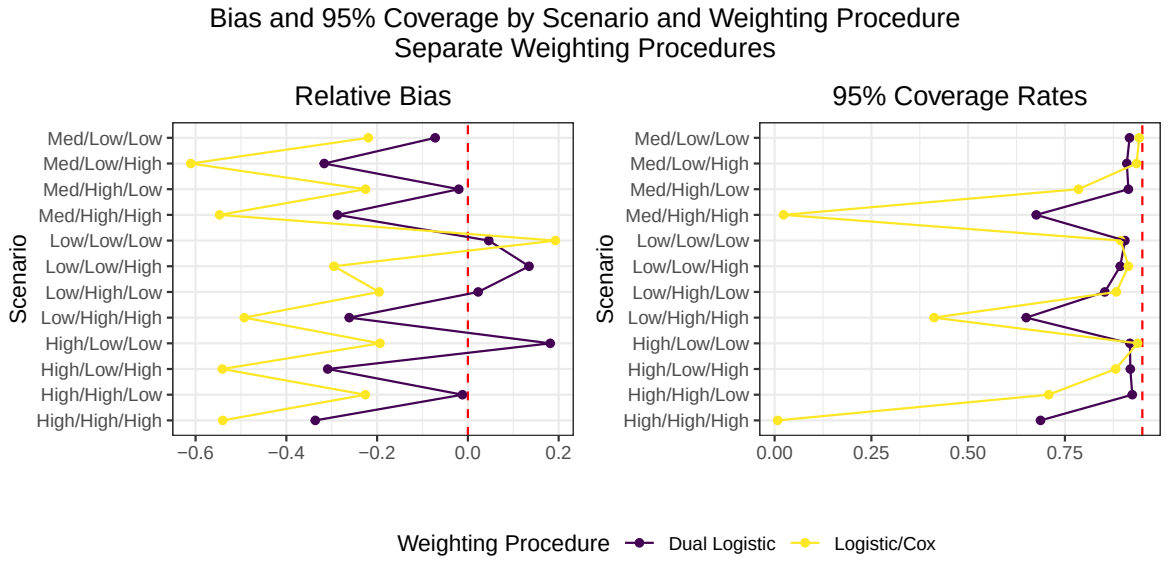
The summarized results for the estimation of the causal estimand are summarized in the following table and plot:

Table 2: Separate Weighting Estimation Results[a]

		Dual Logistic			Logistic/Cox		
Scenario[b]	True ACE	Avg. Estimate	Relative Bias	95% Coverage Rate	Avg. Estimate	Relative Bias	95% Coverage Rate
High/High/Low	0.12	0.1186	-0.0119	0.924	0.0929	-0.2260	0.708
Med/High/Low	0.12	0.1176	-0.0200	0.914	0.0929	-0.2256	0.785
Low/High/Low	0.12	0.1227	0.0227	0.853	0.0965	-0.1956	0.883
High/Low/Low	0.014	0.0165	0.1815	0.918	0.0113	-0.1937	0.938
Med/Low/Low	0.014	0.0130	-0.0720	0.917	0.0109	-0.2194	0.942
Low/Low/Low	0.014	0.0146	0.0460	0.905	0.0167	0.1928	0.894
High/High/High	0.12	0.0797	-0.3362	0.687	0.0552	-0.5401	0.008
Med/High/High	0.12	0.0856	-0.2871	0.676	0.0543	-0.5472	0.023
Low/High/High	0.12	0.0886	-0.2614	0.650	0.0609	-0.4926	0.412
High/Low/High	0.014	0.0097	-0.3088	0.919	0.0064	-0.5410	0.881
Med/Low/High	0.014	0.0096	-0.3165	0.910	0.0055	-0.6106	0.935
Low/Low/High	0.014	0.0159	0.1348	0.893	0.0099	-0.2950	0.914

[a] Obtained over 1000 total bootstrapped replicates (100 bootstrap replicates over each of 100 samples)

[b] Scenarios are described by (degree of overlap/treatment effect level/censoring level)



When using logistic regression to estimate both IPTW and IPCW, we observed smaller relative bias and higher 95% coverage rate among low censoring scenarios regardless of overlap and treatment effect settings.

The high overlap scenarios were also observed associated with higher 95% coverage rate. However, such a pattern was not obvious when using Cox model to estimate IPCW.

4.2 Sequential Weighting

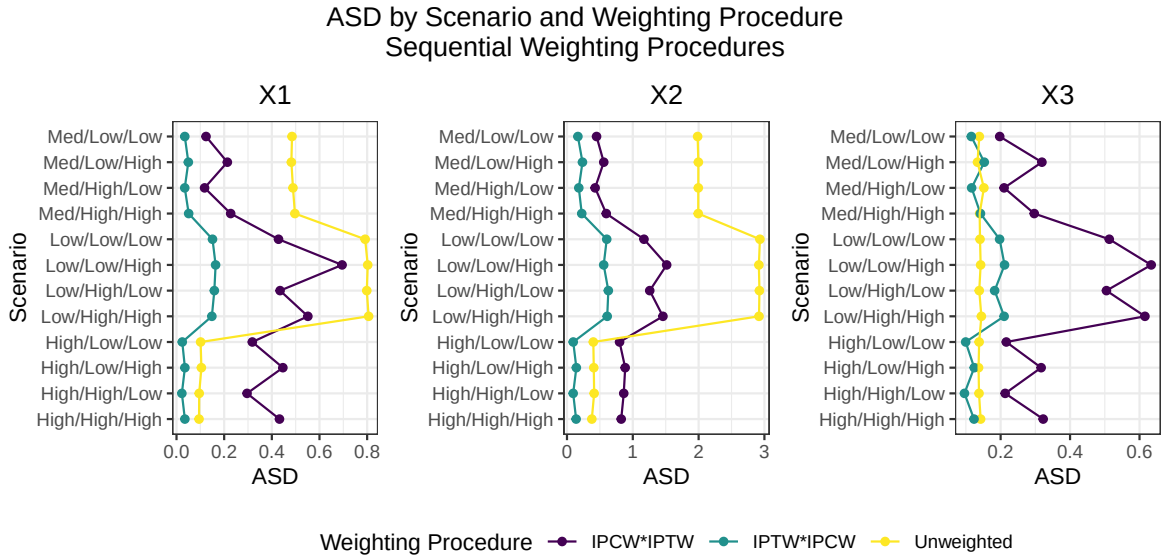
The following table provides results corresponding to how the sequential weighting procedures performed in balancing covariates between the two treatment groups:

Table 3: Sequential Weighting Covariate Balance Results (ASD)[a]

	X1			X2			X3		
Scenario[b]	Unweighted	IPTW∘IPCW	IPCW∘IPTW	Unweighted	IPTW∘IPCW	IPCW∘IPTW	Unweighted	IPTW∘IPCW	IPCW∘IPTW
High/High/Low	0.0952	0.0225	0.2960	0.4127	0.0939	0.8626	0.1378	0.0959	0.2131
Med/High/Low	0.4887	0.0346	0.1178	1.9954	0.1795	0.4275	0.1518	0.1165	0.2100
Low/High/Low	0.7985	0.1588	0.4346	2.9240	0.6283	1.2576	0.1384	0.1828	0.5048
High/Low/Low	0.1011	0.0242	0.3182	0.4002	0.0949	0.7991	0.1381	0.0993	0.2163
Med/Low/Low	0.4846	0.0348	0.1239	1.9852	0.1660	0.4492	0.1387	0.1156	0.1974
Low/Low/Low	0.7916	0.1510	0.4281	2.9317	0.6041	1.1700	0.1407	0.1972	0.5129
High/High/High	0.0945	0.0351	0.4321	0.3772	0.1386	0.8240	0.1425	0.1234	0.3226
Med/High/High	0.4968	0.0510	0.2273	1.9941	0.2244	0.5964	0.1396	0.1419	0.2969
Low/High/High	0.8062	0.1484	0.5512	2.9180	0.6113	1.4578	0.1446	0.2101	0.6156
High/Low/High	0.1044	0.0352	0.4463	0.4069	0.1422	0.8841	0.1365	0.1240	0.3167
Med/Low/High	0.4820	0.0496	0.2134	1.9965	0.2354	0.5590	0.1335	0.1529	0.3189
Low/Low/High	0.8024	0.1641	0.6949	2.9156	0.5579	1.5148	0.1425	0.2113	0.6337

[a] Averaged over 1000 total bootstrapped replicates (100 bootstrap replicates over each of 100 samples)

[b] Scenarios are described by (degree of overlap/treatment effect level/censoring level)



It is apparent that there is improvement across all simulation settings in covariate balance in the IPTW∘IPCW weighting method over the unweighted observed data in covariates X_1 and X_2, though this is not the case in X_3. We note that the covariate balance improved for most cases through the IPCW∘IPTW weighting method over the unweighted observed data in X_1 and X_2, with the exclusion of the high-overlap settings, but the improvement was not as large as the IPTW∘IPCW procedure. In addition, there was greater imbalance in X_3 in the IPCW∘IPTW procedure than both the IPTW∘IPCW and unweighted procedures.

This phenomena may suggest that in terms of covariate balance, the first step of weighting in the sequential weighting procedure may dominate the second set of weights. If this is an accurate conclusion, it is understandable why the covariate balance between treatment groups when first weighting on censoring indication is not as good as the covariate balance between treatment groups when first weighting on treatment, because the distribution of covariates between censoring groups is likely not the same (and in this study is defined to be different) as the distribution of covariates between treatment groups.

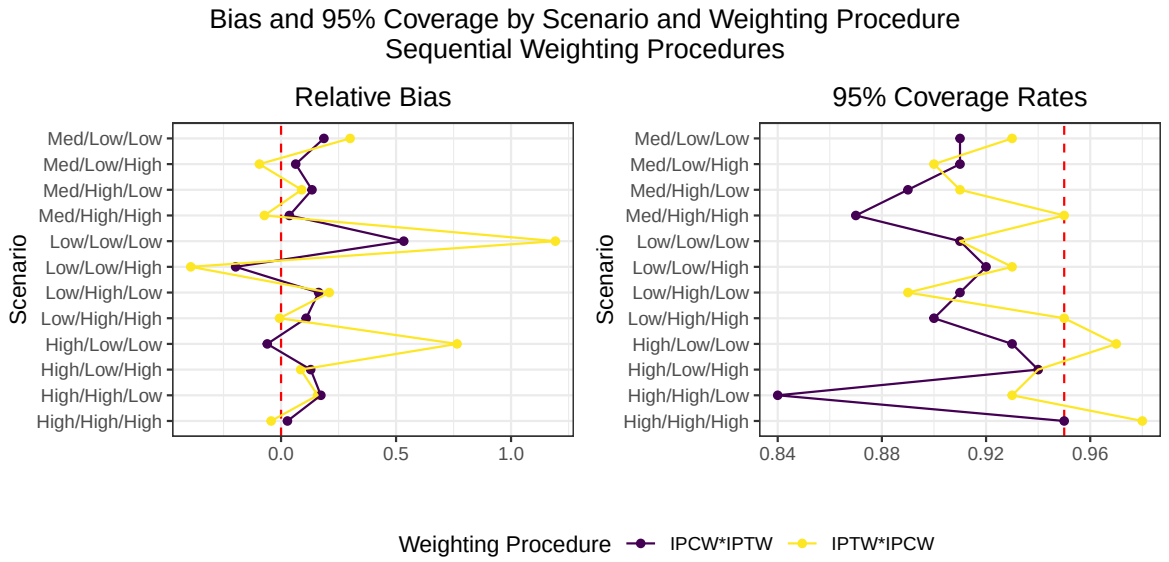
Now, we will assess the results under these two weighting procedures:

Table 4: Sequential Weighting Estimation Results[a]

		IPTW*IPCW			IPCW*IPTW		
Scenario[b]	True ACE	Avg. Estimate	Relative Bias	95% Coverage Rate	Avg. Estimate	Relative Bias	95% Coverage Rate
High/High/Low	0.12	0.1391	0.1590	0.93	0.1408	0.1731	0.84
Med/High/Low	0.12	0.1307	0.0894	0.91	0.1361	0.1345	0.89
Low/High/Low	0.12	0.1452	0.2096	0.89	0.1396	0.1633	0.91
High/Low/Low	0.014	0.0247	0.7653	0.97	0.0132	-0.0600	0.93
Med/Low/Low	0.014	0.0182	0.3003	0.93	0.0166	0.1864	0.91
Low/Low/Low	0.014	0.0307	1.1927	0.91	0.0215	0.5335	0.91
High/High/High	0.12	0.1148	-0.0430	0.98	0.1233	0.0278	0.95
Med/High/High	0.12	0.1113	-0.0724	0.95	0.1244	0.0363	0.87
Low/High/High	0.12	0.1192	-0.0065	0.95	0.1331	0.1092	0.90
High/Low/High	0.014	0.0152	0.0853	0.94	0.0158	0.1288	0.94
Med/Low/High	0.014	0.0127	-0.0939	0.90	0.0149	0.0637	0.91
Low/Low/High	0.014	0.0085	-0.3926	0.93	0.0112	-0.1973	0.92

[a] Obtained over 1000 total bootstrapped replicates (100 bootstrap replicates over each of 100 samples)

[b] Note: scenarios are described by (degree of overlap/treatment effect level/censoring level)



Here, we see that the IPCW∘IPTW seems to produce less relative bias than the IPTW∘IPCW method overall. However, the IPTW∘IPCW method indicates better empirical 95% confidence interval coverage rate than the IPCW∘IPTW method, in which many scenarios had coverage rates that were very close to the aspiration of 95%.

The IPTW∘IPCW method seemed to indicate best performances under settings of high censoring and high treatment effect, irrespective of the degree of treatment overlap. There was less of an observable trend with the IPCW∘IPTW method, but both methods had the best performance in the "all-high" setting, where the relative bias was indicated to be very small and the 95% confidence interval coverage rate was high (95% for the IPCW∘IPTW method and approximately 100% for the IPTW∘IPCW) method.

In these two procedures, high levels of relative bias did not necessarily correspond with lower 95% coverage, and vice versa – in fact, some of the scenarios in which bias for both sequential weighting methods was the lowest also corresponded with lower 95% coverage (specifically the High/High/Low setting), while the scenario that was associated with the highest relative bias for both methods still produced a reasonable 95% coverage in both methods (Low/Low/Low setting).

4.3 Comparison of Procedures

When considering covariate balance, the two separate weighting procedures as well as the IPTW∘IPCW sequential weighting procedure performed similarly. The IPCW∘IPTW procedure produced the most sporadic balancing results out of the four weighting methods.

These weighting procedures produced interesting results with respect to relative bias and 95% confidence interval coverage rates. Both of the sequential weighting procedures seemed to out-perform the separate weighting procedures in the 95% empirical coverage rates, with the lowest coverage in the sequential procedures being around 84% and the majority of which achieving around 92% coverage.

In addition, it was interesting to see that the separate weighting procedures seemed to negatively bias the causal estimates while the sequential procedures seemed to positively bias estimates. The bias in the separate weighting procedures seemed to be more predictable in behavior than the sequential weighting procedures, as low levels of censoring in the separate weighting procedures always corresponded with lower levels of relative bias, holding all other settings constant.

5. Discussion

The results that were obtained through this simulation study were interesting and may compel further inquiry.

The separate weighting procedures involving logistic regressions for both propensity score models performed in ways that matched prior expectations, where scenarios with lower overlap and higher censoring rates scenarios corresponded with lower coverage rates and performance was better overall in lower censoring settings than higher censoring counterparts. However, these results may also prompt further investigation into the inner workings of the `ipcw` function from the `riskRegression` package, as results produced utilizing IPCWs generated from the Cox-Exponential model, were associated with increases in relative bias and decreases in the coverage rates. The sequential weighting procedures both produced 95% coverage rates that were much higher than expected. The covariate balance that was achieved by the IPTW$\circ$IPCW weighting procedure was quite impressive, though it was intriguing that such balance was still associated with large relative biases.

While we are inclined to recommend the sequential IPTW$\circ$IPCW for use on the basis of improved covariate balance and most optimal 95% coverage rates, there are still a number of questions that would necessitate further investigation prior to the employment of such an estimation schematic. First, we recognize the great limitation in our study that a comparative analysis with correctly specified propensity score models was not conducted, and thus the impact of model misspecification on these causal estimates is not yet quantifiable. We hypothesize that a correctly specified propensity score model would decrease the relative

biases that were observed in all scenarios, as it was known that one of the covariates included in every propensity score model was explicitly defined to have no prognostic influence on the indicator of interest. Additionally, due to computational limitations, it may be that the number of simulations that were run were still not sufficient in approximating the long-run behavior of any of these estimators – it would be our recommendation to increase both the sample size and simulation replicate size in subsequent studies to obtain a more complete understanding of the behavior of these procedures and estimators. In the conduct of future simulation studies, we would also recommend varying and adding granularity to the scenario parameters that we used. Finally, as the focus of this study was the restricted average causal estimand, which was estimated by the non-parametric KM estimator, we would recommend investigation into the behavior, or even the application of, these weighting procedures in the setting of other causal estimands and estimators.

6. Github Repository

All code for this study can be found at the following GitHub repository: https://github.com/waveley/BIS537-Final.

7. References

Mao, H., Li, L., Yang, W., & Shen, Y. (2018). On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. Statistics in medicine, 37(26), 3745-3763.

Cheng, C., Li, F., Thomas, L. E., & Li, F. (2022). Addressing extreme propensity scores in estimating counterfactual survival functions via the overlap weights. American journal of epidemiology, 191(6), 1140-1151.

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology (Cambridge, Mass.), 11(5), 550–560. https://doi.org/10.1097/00001648-200009000-00011

Andersen, P. K., & Perme, M. P. (2010). Pseudo-observations in survival analysis. Statistical Methods in Medical Research, 19(1), 71-99. 10.1177/0962280209105020

Zeng, S., Li, F., & Hu, L. (2021). Propensity score weighting analysis of survival outcomes using pseudo-observations. arXiv preprint arXiv:2103.00605.

Wang, C., Wei, K., Huang, C., Yu, Y., & Qin, G. (2023). Multiply robust estimator for the difference in survival functions using pseudo-observations. BMC medical research methodology, 23(1), 247. https://doi.org/10.1186/s12874-023-02065-6

Bender R, Augustin T and Blettner M. (2005). Generating survival times to simulate cox proportional hazards models. Statistics in Medicine; 24: 1713–1723. https://doi.org/10.1002/sim.2059.