JOURNAL OF BIOPHARMACEUTICAL STATISTICS
2022, VOL. 32, NO. 6, 897–914
https://doi.org/10.1080/10543406.2022.2058524

Taylor & Francis
Taylor & Francis Group

Bias and Type I error Control in Correcting Treatment Effect for Treatment Switching Using Marginal Structural Models in Phase III Oncology Trials

Jing Xu, Guohui Liu, and Bingxia Wang

Statistical and Quantitative Science Department, Takeda Pharmaceutical Company, Ltd. Cambridge, Massachusetts, USA

ABSTRACT
This research focuses on the bias and type I error control issues when the marginal structural models (MSMs) are applied to evaluate the causal survival benefits of active intervention versus control in randomized clinical trials (RCTs) with treatment switching after disease progression. When MSMs are applied in the RCT setting, the question of interest, model specifications, strategies for type I error control, bias reduction, etc. differ somewhat from those for observational studies. This manuscript discusses the approaches used to accommodate these differences. Through Monte Carlo simulations and a case study, our research demonstrates that, with sufficient attention paid to issues applicable to RCTs in particular, MSMs may perform better than the inverse probability of censoring weighting (IPCW) method in analyzing the survival endpoint in RCTs with treatment switching because more information is used by the MSM.

ARTICLE HISTORY
Received 18 October 2020
Accepted 18 March 2022

KEYWORDS
Treatment switching; alternative therapy; inverse probability weighting; marginal structure model; time-dependent confounding

1. Introduction

In phase 3 oncology randomized controlled trials (RCTs), patients in both arms are permitted to take alternative treatments after disease progression (hereafter referred to as “treatment switching”) because of ethical considerations. In such situations, although analysis of progression free survival (PFS) is not affected, the effect of active intervention on overall survival (OS) is no longer directly observable. The intent-to-treat (ITT) analysis of the observed data may underestimate the active intervention benefit on OS that would have been observed without treatment switching (Watkins et al., 2013).

To adjust for the time-dependent confounding caused by treatment switching, the inverse probability of censoring weighting (IPCW) (Rimawi and Hilsenbeck 2012; Robins and Finkelstein 2000; Rotnitzky and Robins 2005) method is a popular approach (Watkins et al., 2013). Another inverse probability-based method, marginal structural models (MSMs) (Robins, 2000), has been widely used to analyze confounded data from observational studies (Cole and Hernán 2008; Hernán et al. 2000, 2001; Robins et al. 2000), but the application of MSMs in RCTs remains relatively limited (Farmer et al., 2018). MSMs, according to some statisticians, “tend to be used in observational studies” and “should not be necessary in good quality large RCTs” (Watkins et al., 2013).

Farmer et al. recently conducted a systematic review (Farmer et al., 2018) of publications that used causal inference methodology in analyzing RCT data from 1986 to 2014 and concluded that further efforts are needed to promote the use of MSM and other causal methods within RCTs to maximize their value. Our research and application experience suggested that MSMs, if used appropriately, could

CONTACT Jing Xu jingboston@hotmail.com SQS Department,Takeda Pharmaceuticals, Inc. 35 Landsdowne Street, 02139 Cambridge, Massachusetts, USA.

be a more powerful tool than IPCW in analyzing confounded survival data from oncology RCTs involving treatment switching. To promote awareness of and facilitate the use of MSM in RCTs, we discuss here the properties of the marginal structural Cox proportional hazard models (hereafter also referred to as "MSMs"), when applied in RCT settings.

Here we focus our investigation on obtaining valid causal inference of treatment effect via applying MSMs in RCTs. In Section 2, we compare and discuss differences in MSM settings between observational cohort studies and RCTs. In Section 3, we discuss using the weighted non-parametric Kaplan-Meier estimator by (Xie and Liu, 2005) in the MSM for OS analysis. In Section 4, we compare MSM performance with ITT and IPCW approaches via Monte Carlo simulations in the RCT setting. In Section 5, we discuss measures of bias reduction when MSMs are applied in analyzing RCT data. Finally, we re-analyze Takeda ELM-PC4 phase 3 trial data as a case study in Section 6, followed by discussion and conclusions in Section 7.

2. Differences in marginal structurl model settings between observational cohort studies and randoized clinical trials

The example cohort study we use here is a prospective observational study of the effect of aspirin intake on mortality among patients with breast cancer (Holmes et al., 2010), where time-varying indicators of disease severity (beyond stage at diagnosis) may be both risk factors for OS and determinants of changes in aspirin use during follow-up, i.e., they are time-dependent confounders. The question of interest is the effectiveness of time varying aspirin intake on OS after adjusting for both baseline and time-dependent confounders.

MSMs can be used for such cohort studies to estimate the causal effect of a time-varying exposure in the presence of both fixed and time-dependent confounders using the inverse probability-of-treatment weighted (IPTW) estimators (Hernán et al. 2000; Holmes et al. 2010; Robins et al. 2000), as summarized by the following model:

$$\lambda_T(t|\bar{A}(t), V) = \lambda_0(t)e^{\{\gamma_1\bar{A}(t)+\gamma_2 V\}}, \tag{1}$$

where A(t) represents patient treatment status at time t, V is a vector of baseline covariates, and the overbar represents covariate history. For example, $\bar{A}(t) = \{A(u); 0 \leq u < t\}$ represents patients' treatment history up to time t.

In the RCT setting, A(t) represents the alternative therapy, which is usually not initiated until disease progression, i.e., time-varying; γ for A(t) effect becomes a nuisance parameter, as the question of interest becomes whether the randomized active treatment has a survival benefit over the control. To align with the RCT's primary focus, parameters in equation (1) can be rearranged as

$$\lambda_T(t|\bar{A}(t), V) = \lambda_0(t)e^{\{\gamma_1 V+\gamma_2\bar{A}(t)\}} \tag{2}$$

where the randomized treatment, R, is embedded in the baseline covariate vector, V. In a phase 3 RCT, if the randomization is stratified, the stratification factor is often included as part of V as well because it could improve the precision of the model. To focus our discussion on the R effect, we single out R from V and re-parameterize equation (2) as follows:

$$\lambda_T(t|\bar{A}(t), R) = \lambda_0(t)e^{\{\gamma_0 V+\gamma_1 R+\gamma_2\bar{A}(t)\}}. \tag{3}$$

We assume that, in model (3), patients remain on A(t) once they start it and that the hazard of death at time t depends on a patient's alternative therapy history through the current value at time t. In Section 4, we will use Monte Carlo simulations to evaluate the impact of dropping baseline covariates in the RCT setting.

When treatment switching is allowed, there are time-dependent confounders, $L(t)$, such as disease progression status and laboratory parameters that satisfy the following conditions: 1) they are risk factors for death and also predict treatment switching, and 2) past alternative therapy history modifies the subsequent level of covariates and clinical outcome, such as laboratory parameters and time to death.

In the presence of $L(t)$ satisfying these two conditions, the estimate $\hat{\gamma}_2$ obtained by maximizing the Cox partial likelihood is an (asymptotically) unbiased estimate of the association parameter γ_2 but is a biased estimate of the causal effect of alternative therapy on OS, even if the potential time-dependent confounders $L(t)$ are included in model (3) (Hernán et al. 2000; Robins et al. 2000). The same is true for $\hat{\gamma}_1$ as the estimate for γ_1 in model (3).

To reduce or eliminate bias, we need to fit time-dependent Cox model (3) with the contribution of a patient i to a risk-set calculation performed at time t weighted by the stabilized weights (SWs): $sw_i(t) = sw_i^T(t)*sw_i^C(t)$, as structured by Hernan and Robins (Hernán et al. 2000; Robins et al. 2000), where $sw_i^T(t)$ is the stabilized weight for treatment, defined as follows:

$$sw_i^T(t) = \prod_{j=1}^{int(t)} \frac{P[A(j) = a_i(j)|\bar{A}(j-1) = \bar{a}_i(j-1), V = v_i]}{P[A(j) = a_i(j)|\bar{A}(j-1) = \bar{a}_i(j-1), \bar{L}(j) = \bar{l}_i(j), V = v_i]} \quad (4)$$

and $sw_i^C(t)$ is the stabilized censoring weight, defined as follows:

$$sw_i^C(t) = \prod_{j=1}^{int(t)} \frac{P[C(j) = 0\,|\bar{C}(j-1) = 0, \bar{A}(j-1) = \bar{a}_i(j-1), V = v_i]}{P[C(j) = 0\,|\bar{C}(j-1) = 0, \bar{A}(j-1) = \bar{a}_i(j-1), \bar{L}(j-1) = \bar{l}_i(j-1), V = v_i]} \quad (5)$$

In the expressions (4) and (5), int(t) is the total number of visits up to time t; $\bar{A}(j)$ represents a vector of alternative treatment history up to visit j; $\bar{C}(j)$ represents a vector of censoring history up to visit j; $C_i(t) = 1$ if a patient is right-censored by time t and $C_i(t) = 0$ otherwise; $\bar{L}(j)$ represents the history of the time-dependent confounder up to visit j; and V is the vector of baseline covariates. As pointed out by Latimer et al., 2018, stabilized weights via expression (5) can be regarded as IPCW being used within the MSM to accommodate informative censoring.

After re-weighting, model (3) effectively creates, for a risk set at time t, a pseudo-population in which 1) $\bar{L}(t)$ no longer predicts initiation of alternative therapy at time t – i.e., $\bar{L}(t)$ is no longer time-dependent confounder; 2) the causal association between alternative therapy and survival is the same as in the original study source population (Hernán et al. 2000; Robins et al. 2000); and 3) causal inference of the randomized active treatment effect on survival can be derived from the time-dependent Cox model because time-dependent confounding no longer exists in the pseudo-population at time t. Note that property 3) is new and is unique to the RCT setting. Hence, under the assumption of no unmeasured confounders, given the measured time-dependent risk factor $L(t)$, we can use a conventional time-dependent Cox regression model based on the re-weighted population to make valid inferences about causal effects of both randomized and alternative treatments in RCTs.

Based on the discussion above, as suggested by Yamaguchi and Ohashi (Yamaguchi and Ohashi 2004a, 2004b), to estimate causal effects of active treatment and alternative treatment over control, we define a counter-factual random variable $T_{\bar{A}}$ to be a patient's time of death from randomization if the patient's history of alternative treatment had been $\bar{A}$, rather than the observed history. We can observe $T_{\bar{A}}$ for whose observed alternative treatment history until the observed time of death, agreed with $\bar{A}$, i.e., $T_{\bar{A}} = T$. We consider the following marginal structural Cox proportional hazards model:

$$\lambda_{T_{\bar{A}}}(t|R; \bar{A}(t)) = \lambda_0(t)e^{\{\beta_0 V + \beta_1 R + \beta_2 \bar{A}(t)\}}, \quad (6)$$

where $\lambda_{T_{\bar{A}}}(t|\cdot)$ is the hazard among patients with randomized treatment arm indicator R, in the trial source population, had all patients followed the alternative treatment $\bar{A}$ through time t and were never censored. This is the structural model for the marginal distribution of the counter-factual variable $T_{\bar{A}}$.

Based on the model (6), $\exp(\beta_1)$ has a causal interpretation as a hazard ratio at any time t, had all patients been randomized to the active treatment arm, compared with the hazard had all patients been randomized to the control arm, given that the alternative treatment experience had been the same for all the patients. Similarly, $\exp(\beta_2)$ has a causal interpretation as a hazard ratio at any time t, had all patients received alternative treatment, compared with the hazard had all patients not received alternative treatment, given that all patients had been randomized to the same arm (Yamaguchi and Ohashi 2004a, 2004b).

Model (6) implies a strong underlying assumption: there is no interaction between active treatment and alternative treatment. In general, this assumption may not be true. For example, 1) if active treatment is more effective than the control, patients in the active arm who take alternative treatment after disease progression have experienced one extra line of intervention and thus are more treatment experienced and may benefit less from alternative treatment than the control; 2) on the other hand, if there is a synergistic effect of taking an active treatment followed by alternative treatment, patients in the active arm may benefit more than the control; and 3) it is also possible that there is no interaction between active treatment and alternative treatment. Therefore, unlike for observational studies, for RCTs, the default MSM setting needs to include the R by A(t) interaction term so the relationship between R and A(t) will be data driven, as follows:

$$\lambda_{T_{\bar{A}}}(t|R;\bar{A}(t)) = \lambda_0(t)e^{\{\beta_0 V+\beta_1 R+\beta_2\bar{A}(t)+\beta_3 R\bar{A}(t)\}}. \tag{7}$$

If an interaction exists between R and A(t), with model (7), $\exp(\beta_1)$ has a causal interpretation as a hazard ratio at any time t, had patients been randomized to the active arm, compared with the hazard had patients been randomized to the control arm, under the condition that no patients had received alternative treatment; and $\exp(\beta_1+\beta_3)$ has a causal interpretation as a hazard ratio at any time t, had patients been randomized to the active arm, compared with the hazard had patients been randomized to the control arm, under the condition that all patients had received alternative treatment. Note that the interpretation of $\exp(\beta_1)$ in model (7) is applicable to a different pseudo-population than is model (6). If $\hat{\beta}_3$ is not significantly different from 0, model (7) reduces to model (6).

3. Non-parametric estimation and inference on survival data using the marginal structural model

To compare randomized treatment effects after controlling for treatment switching, adjusted survival curves need to be presented, which can be accomplished by using the baseline option of the PROC PHREG procedure in SAS (SAS Institute INc., 2011), based on the semi-parametric Cox model, as suggested by Cole and Hernan (Cole and Hernan, 2004).

In RCTs, the proportional hazard assumption often does not hold well when OS data is compared between the active and control arms, where non-parametric statistics should be used. To present non-parametric summary from analyzing an OS endpoint in an RCT with treatment switching, we recommend applying the adjusted Kaplan-Meier estimator (AKME) and weighted log-rank test developed by Xie and Liu (Xie and Liu, 2005) under the MSM setting. To implement AKME in MSM,

- First, re-order deaths at D distinct times $t_1 < t_2 < \ldots < t_D$ in the whole sample, at time t_j, j $= 1, \ldots\ldots, D$; ties are allowed. Then, at time t_j, there are d_{jk} deaths out of n_{jk} patients at risk in the treatment arm k.
- We can write weighted version of d_{jk} and n_{jk} as $d^w_{jk} = \sum_{i:T_i=t_j} w_{ik}\delta_i I(X_i = k)$, and $n^w_{ik} = \sum_{i:T_i=t_j} w_{ik} I(X_i = k)$, where i and k are the patient and the treatment arm index, respectively; w_{ik} is the stabilized weight derived in Section 2 for the i^{th} patient in the treatment arm k; δ_i is a survival indicator; $I(X_i = k)$ is the indicator function that equals 1 if patient i belongs to treatment arm k and 0 otherwise.

- The AKME for the treatment arm k is given by:

$$\hat{S}^k(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_j \le t} \left[1 - \frac{d^w_{jk}}{n^w_{jk}} \right] & \text{if} t_1 \le t \end{cases}$$

Xie and Liu proved that $\hat{S}^k(t)$ is an unbiased estimator in the range of the observed data; the variance of $\hat{S}^k(t)$ is in closed form, which is very similar to Greenwood's formula (Kalbfleisch and Prentice 2011).

A two-sample weighted log-rank test used in MSM can be derived based on Xie and Liu's method (Xie and Liu, 2005), as follows:

- Let $d^w_j = d^w_{j0} + d^w_{j1}$ and $n^w_j = n^w_{j0} + n^w_{j1}$ denote the weighted number of deaths and number at risk in the combined sample at time t_j, where $k = 0$ and 1 stands for the control and active treatment arms, respectively; the test of $H_0 : S^1(t) = S^0(t)$ vs. $H_1 : S^1(t) \neq S^0(t) \quad t \le \tau$ will be based on:

$$G^w = \sum_{j=1}^{D} \left(d^w_{j1} - n^w_{j1} \left(\frac{d^w}{n^w_j} \right) \right)$$

The variance of G^w is given by

$$Var(G^w) = \sum_{j=1}^{D} \left(\frac{d_j \times (n_j - d_j)}{n_j \times (n_j - 1)} \sum_{i=1}^{n_j} \left[\left(\frac{n^w_{j0}}{n^w_j}\right)^2 w_i^2 X_i + \left(\frac{n^w_{j1}}{n^w_j}\right)^2 w_i^2 (1 - X_i) \right] \right),$$

where $X_i =$ 0 or 1 (randomized treatment arm index for each individual), and the weighted log-rank test is given by (under Ho)

$$Z_0 = G^w / \sqrt{Var(G^w)} \rightarrow N(0, 1)$$

Simulation results showed that AKME provides a better estimator, closer to the true survival function when the proportional hazards assumption is invalid (Xie and Liu, 2005). Based on the above formulas, SAS (SAS Institute Inc. 2011) macros have been developed at Takeda covering AKME and weighted log-rank test for the adjusted survival analysis using either MSM or IPCW. The implementation needs to use counting process style of data input, which has not been developed by the SAS LIFETEST procedure (SAS Institute Inc., 2011) so far.

4. Simulation studies: algorithm and results

To verify the MSM properties discussed in Section 2 in the RCT setting, Monte Carlo simulations were conducted to test the following three scenarios: 1) MSM performance when there is no interaction between randomized (R) and alternative (A(t)) treatment; 2) MSM performance when the effect of R is modified by A(t); and 3) MSM performance when baseline covariate distributions are not balanced between randomized arms.

4.1. Simulation algorithm

The simulation algorithm is adapted from the approach by (Young et al. 2014) in the observational study setting, with RCT elements being added to the MSM, such as R and the R by A(t) interaction term. Specifically, for patient i at cycle m (assuming a 30-day cycle):

(1) At baseline, $m = 0$, and the randomized treatment arm $R = 0, 1$:
 * generate baseline covariates $x_{1,i} \sim N(\mu_R, \sigma^2)$, and $x_{2,i} \sim Bin(p_R)$;
 * generate survival time under the randomized treatment $T_{R,i} \sim exp(\lambda_0)$, without covariate effects.

* with covariate effects added, $T_i = 30 \times \left(T_{0,i}(1-R) + T_{1,i}R + a{*}x_{1,i} + b{*}x_{2,i}\right)$ is the survival time without A(t), where x_1 and x_2 are the baseline covariates distributed as defined above; a and b are constants; define $L_0 = A_0 = Y_0 = 0$, and $\beta_1 = log\left(\frac{\lambda_1}{\lambda_0}\right) = log(HR)_R$.

(2) Post baseline, $m = 1, 2, \ldots, K$ (assuming study ends after K cycles): generate the time-dependent confounder $L_{m,i}$ from some choice of $f(L_m|\bar{A}_{m-1}, \bar{L}_{m-1}, X_1, X_2, R, Y_m = 0|\theta)$ [18], where $\theta's$ are the regression parameters in the regression model; in this research, we choose f(.) as a linear function of $(\bar{A}_{m-1}, \bar{L}_{m-1}, X_1, X_2, R_i)$: i.e., regard $L_{m,i}$ as a time varying bio-marker that is affected by the past treatment and covariate history.

(3) Assume A(t) is initiated right after disease progression (PD). Generate A(t) $\sim$ bin (p_a), with $p_a =$ Sigmoid(logit $(P(\cdot))$, and $P(\cdot) = logit[P(A_m = 1|\bar{A}_{m-1}, \bar{L}_m, Y_m = 0)] = \alpha_0 + \alpha_1\bar{L}_{m,i} + \alpha_2\bar{A}_{m-1,i}$. The higher the past and current L values, the higher probability Pr(A(t)) = Pr(PD). Thus, R impacts A(t) through $\bar{L}$, and A(t) is determined by past treatment and covariate history.

(4) Get Y_m evaluated at previously generated $(\bar{A}_{m-1,i}, \bar{L}_{m-1,i})$:

* If $T_i > (m-1) \times 30$, then $Y_{m,i} = 0$, patient i is still alive by cycle $m < K$, and then
* If $A_{m-1,i} = 1$, update survival time $T_i = (T_i - 30{*}(m-1))e^{-\beta_2 A_{m-1,i}} + 30(m-1) + c\bar{L}_{m-1,i}$, where β_2 is log(HR) $_A$ versus control and c is constant.
* Go to $m+1$ iteration for patient i;
* Otherwise, $Y_{m,i} = 1$, i.e., patient i died by cycle m, with event time T_i;
* Clearly, survival time is also impacted by the treatment experience (R and A(t)) and covariate history

(5) If $T_i > (K{*}30)$, patient i is administratively censored; separately, generate independent censoring via uniform distribution at a 5% rate while $T_i \leq (K{*}30)$

(6) Finally, to introduce the R by A(t) interaction, for the case A(t) is effective $(\beta_2 = log(HR)_A < 0)$ and patients in the control arm $(R = 0)$ benefited more from A(t)–in step 4 above, update the survival time $T_i = (T_i - 30{*}(m-1))e^{-\beta_2\left(A_{m-1,i}\right)(1-R_i)} + 30(m-1) + c\bar{L}_{m-1,i}$

The simulated clinical trial data were analyzed using extended Cox models in the time-dependent covariate setting via a counting process style data input (SAS Institute Inc, 2011), where no weight was applied for the ITT approach. Treatment switching or censoring probabilities were obtained by Cox regression approximated via pooled logistic regression models (D'Agostino et al., 1990), from which SWs for the MSM were calculated per expressions (4) and (5), respectively. For the IPCW, SW was computed similarly but based on inverse probability of censoring only. Robust variance was used in the final MSM and IPCW models. For the IPCW approach, randomized treatment R was included in the models; after A(t) is initiated, data were censored. For the MSM approach, R and A(t) status were included in the models before and after treatment switching; the R by A(t) interaction was also included in the model when effect modification was generated by the simulations. ITT analysis did not include A(t) in the final model when the simulated data were analyzed.

4.2. Simulation results

4.2.1. MSM performance when there is no interaction between randomized and alternative treatment

If there is no interaction between R and A(t), exp(β_1) from the ITT, IPCW, and MSM approaches are all related to the hazard ratio comparing active with control, where the performance of these three approaches can be compared directly. Table 1, Figure 1 and Figure 2 below summarize simulation results under H_0 and H_1, respectively, where the randomization ratio is 1:1 in each simulation interaction, with baseline covariates generated from same distributions in both the active and control arms.

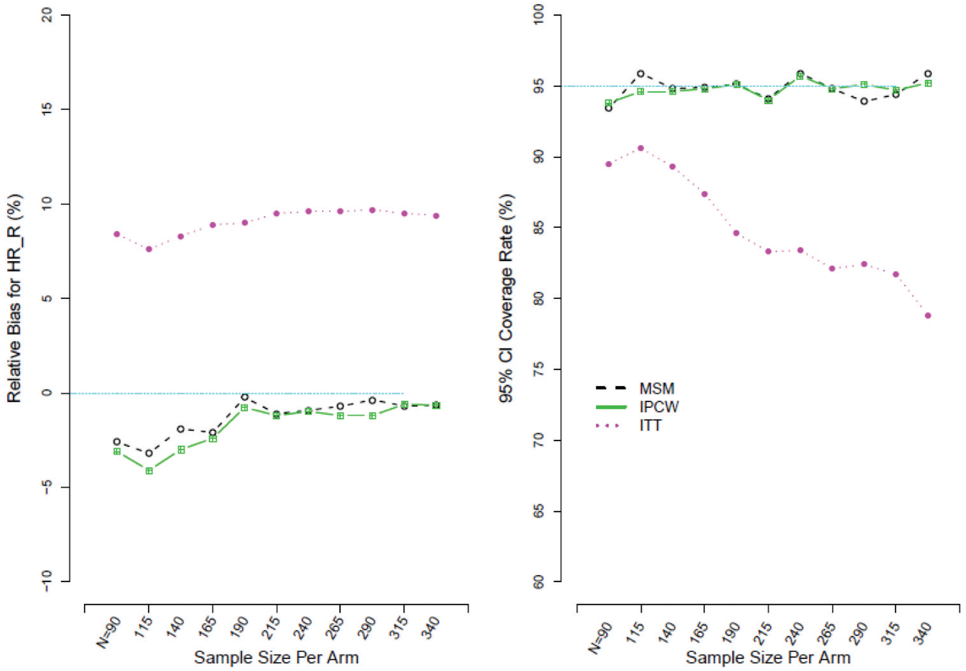

Figure 1. Relative bias and 95 % CI coverage for HR_R, under H_1, ($\lambda_0 = 0.14, HR_R = 0.698, HR_A = 0.741$); Simulation Iterations = 1000.

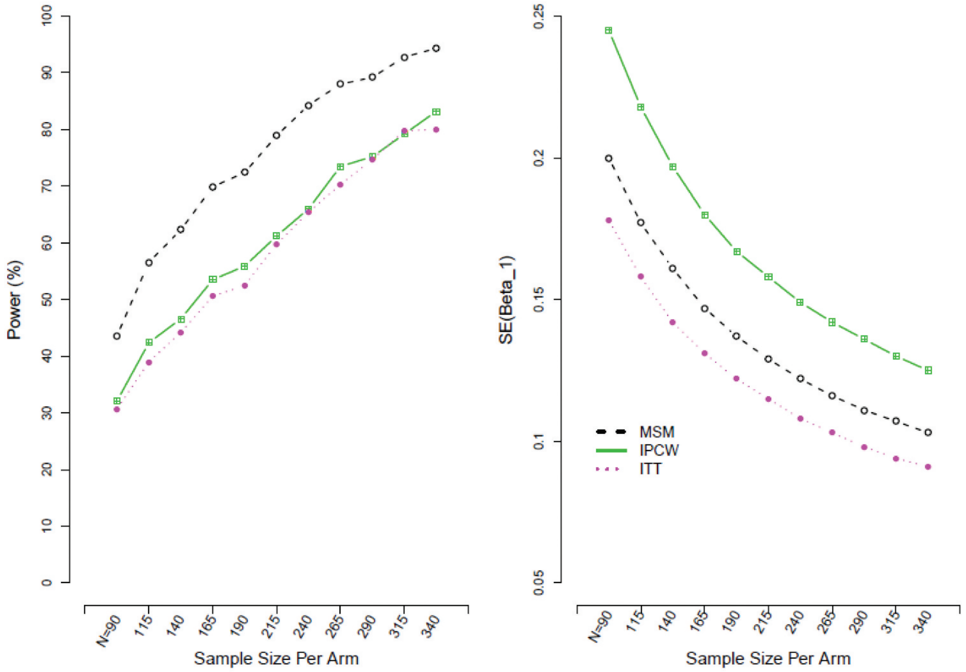

Figure 2. Power for HR_R and SE($\hat{\beta}_1$), Under H_1, ($\lambda_0 = 0.14, HR_R = 0.698, HR_A = 0.741$), Simulation Iterations = 1000.

As shown in Figure 1, for the tested cases, under the alternative hypothesis, for both MSM and IPCW approaches, the relative bias ranges from −1% to −4% when the sample size is less than 200 per arm and is less than −1% when the sample size is more than 200 per arm; however, the relative bias is consistently above 8% for the ITT approach for all tested cases, indicating that the hazard ratio is likely to be inflated by the ITT approach when treatment switching occurs in an RCT. The 95% CI coverage rates are close to the nominal levels for all sample sizes for the MSM and IPCW approaches, but the coverage rate is very poor for the ITT approach, which also deteriorates with the increase of sample sizes. As shown in Figure 2, if model assumptions are not violated, since the MSM uses all observed data, it yields a smaller standard error (SE) for β_1 than the IPCW. SEs for β_1 from both IPCW and the MSM are larger than those from the ITT approach because additional variability is introduced by the inverse probability weighting. Power is lower for the IPCW approach than to the MSM because of the larger SEs due to information loss by censoring data after treatment switching. For the ITT analysis, power is also lower than that for the MSM, given the inflated hazard ratio estimates, as SEs for β_1 from the ITT approach are smaller than those from the MSM.

As shown in Table 1, if model assumptions hold for the tested cases, under the null hypothesis, with moderate to large sample sizes, the MSM performs well for inference on the randomized treatment effect, with type I error controlled at the nominal level. The bias and standard error for β_1 are in the same range as those from the ITT approach and both are smaller than those from the IPCW approach. In addition, under the null hypothesis, the MSM also performs well for inference on the alternative therapy treatment effect, which could not be accomplished by either the ITT or IPCW approach.

4.2.2. MSM performance when there is interaction between randomized and alternative treatments

As discussed in Section 2, when there exists interaction between R and A(t), if the interaction term is omitted in MSM equation (7), it is a model misspecification, the impact of which is summarized in this section via simulation.

To simulate the R by A(t) interaction, data are generated using the formula shown in item 6 in Section 4.1, where the control arm will benefit more from A(t) than the active arm after A(t) is initiated. If we change $(1 - R_i)$ to $(1 + R_i)$ in the same formula, then the active arm will benefit more from A(t) than the control arm. By this approach, when data are generated under the null hypothesis, $\beta_1 = \beta_2 = 0$, it will incur no interaction between randomized and alternative

Table 1. MSM Performance When There is no Interaction Between R and A(t), Under $H_0 : \beta_1 = 0, \beta_2 = 0$.

Simulation Parameters: $\lambda_0 = \lambda_1 = 0.14$, HR_R =1; Simulation iterations = 10,000							
Sample Size per Arm	Model Type	Observed Mean HR_R	Absolute Bias	Relative Bias (%)	Type I Error (α)	95% CI Coverage	SE ($\hat{\beta}_1$)
100	ITT	1.0157	0.0157	1.57	0.0488	0.9512	0.1699
	IPCW	1.0268	0.0268	2.68	0.0497	0.9503	0.2321
	MSM	1.0161	0.0161	1.61	0.0505	0.9495	0.1701
200	ITT	1.0072	0.0072	0.72	0.0498	0.9502	0.1198
	IPCW	1.0131	0.0131	1.31	0.0513	0.9487	0.1637
	MSM	1.0073	0.0073	0.73	0.0501	0.9499	0.1200
300	ITT	1.0045	0.0045	0.45	0.0486	0.9514	0.0977
	IPCW	1.0075	0.0075	0.75	0.0472	0.9528	0.1335
	MSM	1.0044	0.0044	0.44	0.0500	0.9500	0.0978
500	ITT	1.0037	0.0037	0.37	0.0495	0.9505	0.0756
	IPCW	1.0059	0.0059	0.59	0.0504	0.9496	0.1033
	MSM	1.0037	0.0037	0.37	0.0504	0.9496	0.0758
Simulation Parameters: $HR_A = 1$; Simulation iterations = 10,000							
Sample Size per Arm	Model Type	Observed Mean HR_A	Absolute Bias	Relative Bias (%)	Type I Error (α)	95% CI Coverage	SE ($\hat{\beta}_2$)
100	MSM	1.0256	0.0256	2.56	0.0522	0.9478	0.2070
200	MSM	1.0160	0.0160	1.60	0.0506	0.9494	0.1459
300	MSM	1.0131	0.0131	1.31	0.0522	0.9478	0.1190
500	MSM	1.0083	0.0083	0.83	0.0515	0.9485	0.0921

Table 2. The MSM Performance When There is Interaction Between R and A(t), Under : $H_1 : \beta_1 < 0, \beta_2 < 0$

Simulation Parameters: $HR_R = 0.7408, HR_A = 0.7408$; Simulation Iterations = 1000

Interaction Term in MSM?	N per Arm	Observed Mean HR_R	Relative Bias (%)	95% CI Coverage	Observed Mean HR_A	Relative Bias (%)	95% CI Coverage
Control Arm Benefited More From A(t)							
Included	140	0.744	0.44	0.940	0.737	−0.515	0.941
	240	0.745	0.62	0.947	0.740	−0.12	0.947
	340	0.747	0.89	0.941	0.743	0.35	0.943
Omitted	140	0.810	9.29	0.895	0.840	13.38	0.887
	240	0.809	9.23	0.881	0.809	13.1	0.862
	340	0.813	9.79	0.807	0.813	13.1	0.810
Active Arm Benefited More From A(t)							
Included	140	0.746	0.64	0.942	0.749	1.11	0.940
	240	0.742	0.15	0.943	0.743	0.35	0.932
	340	0.745	0.55	0.943	0.742	0.17	0.954
Omitted	140	0.683	−7.80	0.895	0.662	−10.70	0.873
	240	0.682	−7.80	0.871	0.659	−11.05	0.839
	340	0.686	−7.44	0.839	0.660	−10.95	0.813

treatments—i.e., the simulation outcome will be very similar to those summarized in Section 4.2.1. Thus, the simulation results under the null hypothesis will not be presented again in this section.

Table 2. Summarizes the MSM performance when R is modified by A(t) and where data are generated under the alternative hypotheses for both R and A(t). From Table 2, in the presence of the R by A(t) interaction, when the interaction term is included in MSM equation (7), for the simulated cases, the $\widehat{HR}_R$ and $\widehat{HR}_A$ are very close to the true values of HR_R and HR_A, with relative bias within the simulation error range; the 95% CI coverage rates are close to the nominal level. However, if the interaction term is omitted from MSM equation (7), both $\widehat{HR}_R$ and $\widehat{HR}_A$ tend to carry non-ignorable bias, and the 95% CI coverage rates are below 90%. Specifically, when the control arm benefits more from the R by A(t) interaction, $\widehat{HR}_R$ tends to be inflated by around 9%; and $\widehat{HR}_A$ tends to be inflated by around 13%–i.e., both randomized and alternative treatment effects tend to be underestimated. When the active arm benefits more from the R by A(t) interaction, $\widehat{HR}_R$ tends to be deflated by approximately 8%; and $\widehat{HR}_A$ tends to be deflated by approximately 11%–i.e., both randomized and alternative treatment effects tend to be overestimated.

Table 2 summarizes the simulated outcome from the MSM only. The outcomes and properties from the ITT and IPCW analyses on the same simulated datasets are given here without tabulations because of space constraints: 1) under the alternative hypothesis, in the presence of the R by A(t) interaction, results from the ITT analyses are always biased, regardless of whether the R by A(t) interaction term is included in the analyses; and 2) no interaction term needs to be included in the IPCW models, as data will be censored after A(t) initiation. As long as the IPCW model assumptions are not violated, which is true for the data simulated in this section, the $\widehat{\beta_1}$ from the IPCW analysis will be an unbiased estimate of β_1. For causal interpretation of the IPCW results, $\exp(\beta_1)$ represents a hazard ratio in the active arm compared to the control arm, had all patients remained on their randomized treatment (Robins and Finkelstein 2000).

4.2.3. MSM Performance When Baseline Variable Distributions Are not Balanced between Randomized Arms

So far, baseline covariate data have been generated under the same distributions for each randomized arm in our simulations. Results in Sections 4.2.1 and 4.2.2 show that so long as baseline covariate distributions are balanced between randomized arms, MSM performance is not affected by excluding baseline covariates in the final MSMs.

Table 3. MSM Performance When the Baseline Covariate Distributions Are Not Balanced

Baseline Covariate Distribution: $X_{1i} \sim N(15, 6)$, $X_{2i} \sim$ Bin(p =0.3) $\in$ R = 1; and $X_{1i} \sim N(10, 6)$, $X_{2i} \sim$ Bin(p =0.2) $\in$ R = 0								
Simulation iterations: 10,000 and 1,000 under H_0 and H_1, respectively; n = 500 per arm								
Randomized treatment Effect, Under H_0 : $\mathrm{HR}_R = 1$ // Under H_1 : $\mathrm{HR}_R = 0.698$								
Covariates Included in the Model?	Model Type	Observed HR_R	Relative Bias (%)	Type I Error	SE ($\hat{\beta}_1$)	Observed HR_R	Relative Bias (%)	95% CI Coverage
Yes	ITT	1.004	0.46	0.048	0.082	0.765	9.65	0.840
	IPCW	1.016	1.61	0.050	0.141	0.692	−0.88	0.947
	MSM	1.004	0.35	0.049	0.083	0.695	−0.38	0.953
No	ITT	0.971	−2.83	0.071	0.076	0.744	6.67	0.889
	IPCW	0.894	−10.61	0.130	0.154	0.638	−8.49	0.886
	MSM	0.968	−3.20	0.076	0.078	0.671	−3.76	0.928
	Alternative Therapy Effect, Under H_0 : $\mathrm{HR}_A = 1$					// Under H_1 : $\mathrm{HR}_A = 0.741$		
Covariates Included in the Model?	Model Type	Observed HR_A	Relative Bias (%)	Type I Error	SE($\hat{\beta}_2$)	Observed HR_A	Relative Bias (%)	95% CI Coverage
Yes	MSM	0.998	−0.23	0.051	0.12	0.739	0.18	0.945
No	MSM	1.102	10.16	0.087	0.199	0.746	0.72	0.944

To investigate the MSM performance when the baseline covariate distributions are not balanced between randomized arms, the baseline data generating settings in step 1 of Section 4.1 are modified, where two sets of baseline variables, X_{1Ri}, and X_{2Ri}, are generated, with R = 0,1 representing the control and active arms, respectively, and i being the patient index. As shown in Table 3, X_{1Ri} is normally distributed, with a mean difference of 5 units, 10 vs. 15, for the control and active arms, respectively; X_{2Ri} follows a Bernoulli distribution, with mean proportion parameters differing by 10 percentage points, 20% vs. 30% for the control and active arms, respectively. Other simulation settings are the same as illustrated in Section 4.1. Since $L_{m,j}$ is a linear function of X_1 and X_2 per item 2 in Section 4.1, the baseline covariate imbalance will unevenly impact A(t) history and OS outcome in each randomized arm. Simulation cases are generated assuming that no R by A(t) interaction exists.

From the simulation results in Table 3, in an RCT with treatment switching, if baseline covariate distributions are not balanced between randomized arms, but those baseline covariates are included in the final MSM, the model performs well: under the null hypothesis, type I error rates are controlled at the nominal 0.05 level, and relative bias is less than 2%. However, if imbalanced baseline covariates are omitted in the final MSM, under the null hypothesis, type I error may increase to beyond the nominal level for the causal inference of both the randomized and alternative treatment effects; relative bias may also increase several folds. Under the alternative hypothesis, relative bias may increase multiple times, and 95% CI coverage rate may be poor for the randomized treatment effect.

As shown in Table 3, omitting imbalanced baseline covariates in the final IPCW models will negatively impact its performance as well. The direction and magnitude of the negative impact are similar to those for the final MSM. Of note, the ITT approach performs poorly under the alternative hypothesis for inference of the randomized treatment effect in the presence of treatment switching with or without imbalanced baseline covariates included in the models, the relative bias is at least 6% for the estimated hazard ratio between the active and control arms, and 95% CI coverage rates are below 90%. This outcome is consistent with the ITT results in Section 4.2.1.

5. Reducing execution bias when applying the MSM in randomized clinical trials

When MSMs are applied to analyzing data from observational studies, diagnostics and sensitivity analyses are essential for their proper use (Cole and Henan, 2008). The same careful practice needs to be exercised when the MSM is used to analyze data from RCTs with treatment switching, so that bias in causal inference is avoided or reduced. Taking advantage of the controlled setting, most measures could and need to be prespecified in the statistical analysis plan (SAP).

5.1. Accommodating unmeasured confounder and consistency assumptions

The assumption of no unmeasured confounding is not testable. In the RCT setting, to accommodate this assumption adequately, it is important to make sure that a rich collection of prognostic factors be collected at the trial planning stage – i.e., all major potential confounding variables, both baseline and time dependent, that may affect disease progression, receipt of alternative therapy, and or death should be identified in the trial design stage, included in the case report forms and listed in the SAP. This does not mean to include as many variables as possible in the Cox regression approximated by the pooled logistic regression models (D'Agostino et al., 1990) for weight computing because: 1) the addition of too many covariates relative to number of observations may introduce finite sample bias due to non-positivity, which will be discussed in Section 5.2; and 2) if non-confounding variables are also included in the pooled logistic regression models for deriving weights, the efficiency of the effect estimates will be negatively affected (Cole and Hernán 2008). Therefore, the identification of major potential confounding variables needs to be pre-planned carefully, via feedback from clinicians and key opinion leaders (KOL's), and based on intensive literature research. Then, these confounders needed to be prespecified in the SAP as covariate candidates in the initial pooled logistic regression models for creating stabilized inverse treatment or censoring probability weights.

The consistency assumption in causal inference means that a subject's counterfactual outcome under his/her observed exposure history is precisely his/her observed outcome (Cole and Frangakis 2009; Hernán et al. 2001). The consistency assumption is rarely exactly true and is also difficult to verify. However, if an RCT's execution closely follows the study protocol, consistency is a reasonable assumption as one can imagine how to manipulate hypothetically an individual's treatment status (Hernán et al., 2001). Otherwise, the assumption may be grossly violated, and bias will be introduced in the MSM causal inference. For example, if there is widespread of exposure misclassification or non-adherence to randomized treatment in an RCT, the consistency assumption may no longer be plausible. For such cases, ad-hoc sensitivity analyses need to be conducted to investigate departure from the consistency assumption and the associated bias introduced in the MSM causal inference outcome.

5.2. Accommodating the positivity assumption

For RCTs with treatment switching, positivity is the condition that patients have a non-zero probability of receiving or not receiving alternative therapy at every level of the combination of confounder set values. When the positivity assumption is violated, bias will be introduced in the MSM causal inference (Cole and Hernán 2008; Robins et al. 2000). One example of the positivity violation is to apply the MSM, as specified in Section 2, directly to RCTs with a cross-over design, where only patients in the control arm are allowed to switch to the active arm after disease progression. If the MSM equation (6) or (7) is used without modifications for the cross-over trial, the causal inference will be biased because of structural non-positivity.

Besides structural zeros, random zeros will also introduce non-positivity (Cole and Henan, 2008), which usually occurs by chance, with zero proportions exposed or not exposed to alternative therapy for particular exposure and covariate histories. In RCTs, when the MSM is applied to analyze confounded OS data, the likelihood of random zero increases mainly owing to the following three reasons: 1) the trial sample size is small, 2) too many weak confounding variables are included in the pooled logistic regression model for inverse probability weight calculations, or 3) data become sparse near the end of long-term follow-up.

Figure 1 in Section 4.1 provides one example of the negative impact of small trial sample size, where the MSM models are correctly specified and all confounders have been accounted for in the weight calculations in the simulation setting, but the relative bias is still more than 3% when N is less than 150 per arm. This may be explained by the random zero-induced non-positivity, given limited sample size or limited number of events. If N is more than 200 per arm, relative bias decreases and stays near the zero percent line.

To reduce random non-positivity associated with putting too many unnecessary covariates in the pooled logistic regression models for deriving inverse probability weights, it is essential to pre-specify in the SAP the criterion for selecting only covariates that are truly associated with treatment switching or censoring distributions. Priorities should be set according to strengths of (suspected) causal relationship of a covariate with the outcome, as small imbalances in a strong predictor for the outcome may cause more severe (residual) confounding bias than large differences in covariates that are only weakly related to the outcome.

In addition, it is also necessary to prevent individual patients from contributing too much in the causal inference of randomized treatment effects. For this purpose, weight truncation criteria need to be pre-defined in the SAP and strictly followed. For example, reset all extremely large weights randomly to less than 2 times the 99th percentile of the weight distribution.

A necessary condition for well-behaved weight model specification is that the stabilized weights are close to a mean of 1 (Hernán et al., 2001). After considering all the factors discussed above, the final weight model chosen is a bias-variance trade-off between the inclusion of a sufficient number of confounders and the construction of well-behaved model that led to a small variance of the effect estimate. Refer to Hernán et al. 2001 for an insightful example on how to settle on weight models.

5.3. Correct model specifications

As discussed in Section 2 and demonstrated in Section 4.2.2 and 4.2.3 via simulations, substantial relative bias for causal inference of both randomized and alternative treatment effects may occur because of two types of model misspecifications in MSM analysis in the RCT setting: 1) omitting the R by A(t) interaction term without justification from observed data, and 2) omitting baseline confounding factors in the final model when they are not balanced between randomized arms. Therefore, in the SAP, measures to prevent these two types of model misspecifications need to be pre-planned.

Estimated weights with very extreme values or mean weights far from 1 are indicative of misspecification in models for inverse probability weight calculations. If the necessary condition for the correct model specification is violated (Cole and Henan, 2008), it is important to check whether any key time-dependent confounders have been left out of the weight calculation models or whether the weight models are otherwise not correctly specified.

The general MSM specification in Section 2 in the RCT setting is based on the ITT type of assumption for the alternative therapy – i.e., A(t) is continuously used without change after it is initiated by a patient. However, in RCTs, this assumption often may be violated. For some patients, after A(t) is initiated, it may be stopped early or a different type or multiple types of A(t) may be used, which may make the correct causal inference impossible for a particular alternative treatment effect. However, this will not affect the valid causal inference for the randomized treatment effect, which is the primary focus of the MSM application in RCTs.

6. A case study

As a case study, we re-analyzed OS data from the ELM-PC4 RCT (Saad et al., 2015). On May 14, 2014, Takeda announced the results from ELM-PC4, a pivotal, international, double-blind, randomized phase 3 trial in men with metastatic, castration-resistant prostate cancer (mCRPC) who had not received chemotherapy. The RCT showed that Orteronel plus prednisone (hereafter referred to as the "active arm") improved radiographic progression-free survival (rPFS) compared to placebo plus prednisone (hereafter referred to as the "placebo arm"), at the time of final analysis. Hazard ratio (HR) for rPFS = 0.71 (95% CI 0.63–0.80); p < .0001; median rPFS = 13.8 and 8.7 months for the active and placebo arms, respectively (see Figure 2A in Saad et al., 2015). However, the ITT analysis did not demonstrate a statistically significant improvement in the study's second primary endpoint of OS. HR for OS = 0.92 (95%CI 0.79–1.08); p = .31; median OS = 31.4 and 29.5 months for the active and control arms, respectively (see Figure 3A in Saad et al., 2015).

Overall, more placebo patients (n = 395, 51%) received subsequent therapy than the patients in active arm (n = 353, 45%) (Saad et al., 2015). Post hoc analysis of the time to starting alternative therapy indicated that placebo patients initiated alternative therapy significantly earlier than patients in active arm as well, as summarized in Figure 3; HR = 0.82 (95%CI 0.71–0.95); p = .0068 from log-rank test; median time to first alternative therapy = 17.7 and 22.2 months for the control and active arms, respectively.

Among possible reasons, the rPFS gain did not translate into an OS benefit for the active intervention over placebo is switching to effective alternative therapies (Saad et al., 2015). Both MSM and IPCW were not pre-specified in the ELM-PC4 trial SAP as sensitivity analyses to investigate the impact of alternative therapies on the OS outcome. In this case study, we use MSM and IPCW to reanalyze the ELM-PC4 data per the procedures and principles discussed in Sections 2 and 5.

In step 1 of the MSM analysis, Cox proportional hazard regression approximated via the pooled logistic models (D'Agostino et al., 1990), is used to generate the probability of treatment switching or censoring. After going through the exercises suggested in Section 5, five time-dependent covariates (Eastern Cooperative Oncology Group (ECOG) performance status, prostate-specific antigen (PSA)

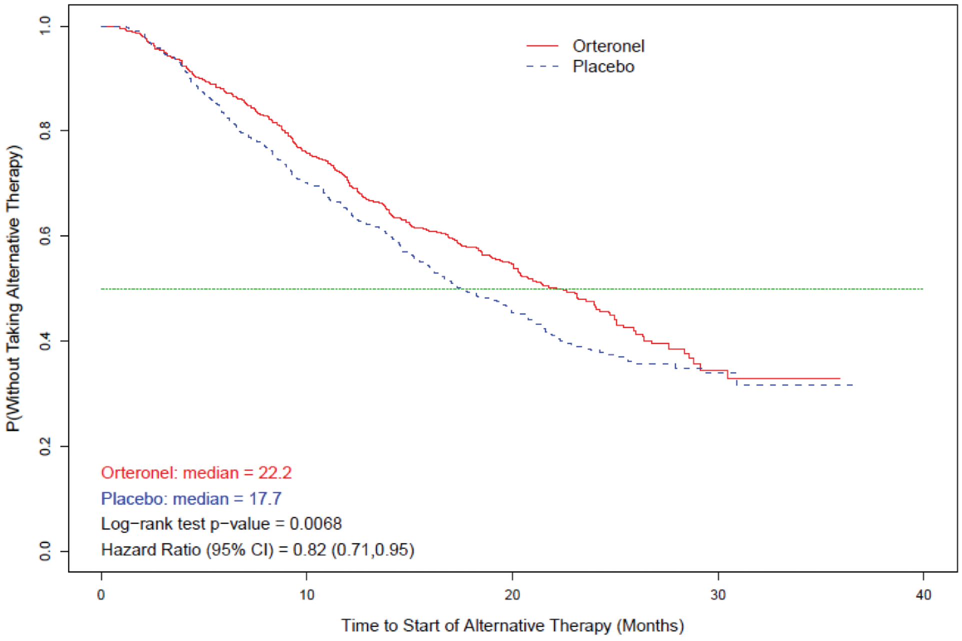

Figure 3. ELM-PC4 Trial: Time to Start of Alternative Therapy.

level, hemoglobin, disease progression status, and study months) and 10 baseline covariates (randomized treatment, region, androgen use at baseline, race, baseline ECOG score, alkaline phosphate category, lactate dehydrogenase (LDH) category, visceral disease status, adrenocorticotropic hormone (ACTH) flag, and prior anti-neoplastic therapy) are kept in the final four models for deriving SWs.

Without limiting the extreme $\widehat{SW}(t)$ values, the overall mean of stabilized weights at all study visits is 0.98, with a standard deviation of 0.46 and range of (0.11, 20.69). Figure 4 summarizes the $\widehat{SW}(t)$ value distributions by study quarter before truncating the extreme $\widehat{SW}(t)$ values. The weight distributions over time look balanced, and there are only a few skewed values at a stabilized weight of approximately 5. After truncating the extreme $\widehat{SW}(t)$ values randomly to 1 to 2 times the 99^{th} percentiles, the overall mean of stabilized weights at all study visits decreases to 0.97, with a standard deviation of 0.33 and range of (0.11, 4.87).

In step 2, the primary MSM analysis is based on the $\widehat{SW}(t)$ after truncation, such that the maximum individual patient contribution at any study visits is less than 5. Since baseline demographics and disease characteristics were balanced between the randomized treatment arms (see Table 1 of Saad et al., 2015), no baseline covariates besides R were included in the final MSM model. The extended weighted Cox regression is stratified by prior anti-androgen therapy (yes, no), region, and radiographic disease progression at baseline, the same as used in the original ITT analysis (Saad et al., 2015). In addition, a sensitivity MSM analysis is conducted without truncating the extreme SW values.

The MSM analysis showed no evidence of interaction between active and alternative treatment for the ELM-PC4 data, as the p-value for the R by A(t) interaction term is 0.5673 in the extended Cox model. On the other hand, the p-value for testing proportionality (Hosmer Jr. et al., 2008) is 0.0606 with degrees of freedom = 2. P-values for log (study time) × R and log (study time) × A(t) terms are 0.4659 and 0.0209, respectively. The result indicates that the proportionality assumption does not hold well for the ELM-PC4 trial data, with A(t) effect being not proportional in the semi-parametric Cox model. Thus, as discussed in Section 3, a non-parametric AKME and a weighted Log-rank test (Xie and Liu, 2005) need to be used to summarize the MSM results.

Figure 5 compares the ELM-PC4 OS outcomes from the ITT analysis and the MSM adjusted analysis. After the MSM adjustment, had the A(t) effect being the same in both randomized arms, the survival probability for the placebo patients would have been lower, and the survival probability for

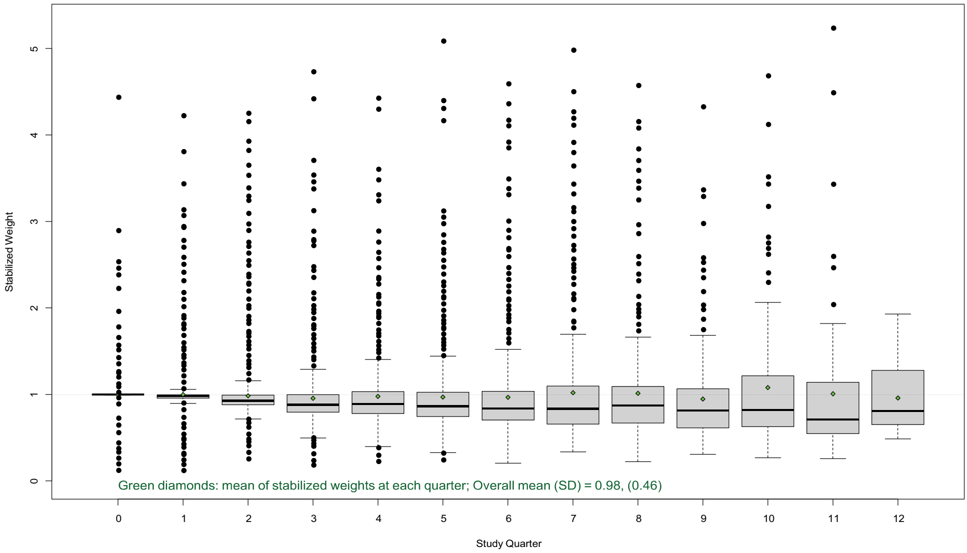

Figure 4. ELM-PC4 Trial: Stabilized Weight Distribution by Study Quarter.

patients in the active arm would have been higher in the late stage of the trial; the median difference between active and control arms would have been 7.5 months (95%CI 34.6–27.1), HR = 0.831, with a p-value of 0.0265 from the adjusted log-rank test and 0.0491 from the robust Wald test from the Cox regression model. From the sensitivity MSM analysis without truncation of extreme SW values (not shown in Figure 5), the median difference would have been 7.0 months (95%CI 34.6–27.6), HR = 0.807, with p-value of 0.0369 from the adjusted log-rank test and 0.0368 from the robust Wald test from the Cox regression model.

As demonstrated by expression (5) in Section 2, and by step 1 of the MSM analysis in this Section, the IPCW is used within the MSM to estimate treatment effect adjusting for informative censoring. In addition, the IPCW is applied as an independent approach in reanalyzing the ELM-PC4 data where both treatment switching and premature loss to follow up are treated as informative censoring, with weights computed via a similar procedure as in step 1 of the MSM analysis. The covariate set for SW computing in the final two models for deriving the censoring probabilities for the IPCW analysis are the same as those used in the MSM analysis. The extended weighted Cox model is stratified the same way as in the original ITT analysis (Saad et al., 2015). AKME and the weighted log-rank test (Xie and Liu, 2005) are also used to summarize the IPCW results. Figure 6 compares the ELM-PC4 OS outcomes from the ITT analysis and the IPCW adjusted analyses. After the IPCW adjustment, the median difference between the active and placebo arms would have been 3 months (95%CI 30.1–27.1), HR = 0.825, with a p-value of 0.0929 from the adjusted log-rank test and 0.0899 from the robust Wald test from the Cox regression model.

Re-analysis of the ELM-PC4 data, after adjusting for the time-dependent confounding introduced by treatment switching, both MSM and IPCW analyses reduced hazard ratio estimate by 10% compared to the ITT analysis, and the p-value from the MSM is smaller than that from the IPCW approach. This outcome is consistent with the Monte Carlo simulation conclusion in Section 4.2.1: the MSM may be more powerful because all observed data are used. The survival curves estimated by AKME in MSM and IPCW look similar in the first half of the study, before most patients initiated alternative therapy; but in the last part of the trial, the adjusted trend was maintained longer with the MSM than with the IPCW. The difference is associated with different $\hat{\beta}_1$ interpretations of the

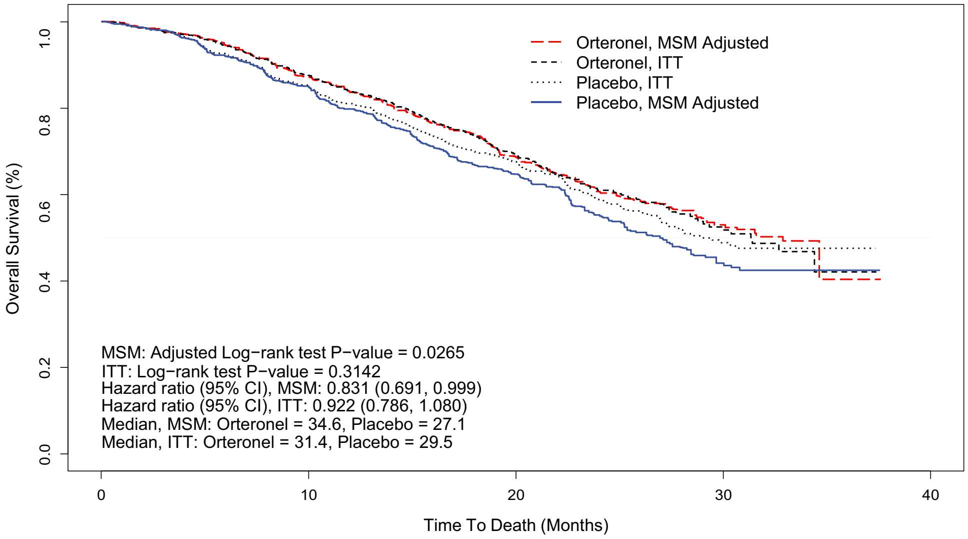

Figure 5. ELM-PC4 Trial: Overall Survival, ITT Analysis vs MSM Adjusted Analysis.

outcomes for this dataset: the adjusted outcome from MSM is applicable to a pseudo-population that would have experienced equal A(t) exposure in both randomized arms; the outcome from IPCW is applicable to a pseudo-population that would have been observed in the absence of switching to A(t). In summary, results from both approaches indicate that the original ITT analysis has underestimated Orteronel's OS benefit because of time-dependent confounding by switching to other effective alternative therapies.

7. Discussion and Conclusion

In this article, we demonstrate that the MSM can be used in analyzing RCT data with treatment switching but with different model specifications than used in the observational studies. Special attention needs to be paid to including the R by A(t) interaction term in the MSM as the default setting. If MSM is used appropriately in analyzing the RCT data, bias and type I error can be well controlled.

The MSM is appealing because it is intuitive and is a natural extension of the time-dependent Cox model. By using AKME and the Cox partial likelihood estimator jointly, outputs from MSM such as the hazard ratio, weighted log-rank test, and adjusted median estimates resemble those from the Cox proportional-hazards regression model and Kaplan-Meier analysis. Thus, it is easy to communicate MSM results to clinicians and non-statisticians.

The IPCW also shares the advantage of easy communication but may be less powerful than the MSM, as information is censored after treatment switching. Applying the MSM jointly with the IPCW to analyze RCT data may provide additional insight about OS distribution with and without treatment switching. When there is an interaction between the randomized and alternative treatments, interpretation of causal inferences on β_1 from the IPCW is similar to that from the MSM: both approaches assume that patients did not receive A(t). For such cases, because of the R by A(t) interaction, survival curves estimated by AKME for the MSM will cross; therefore, median survival for patients who would have remained on the randomized treatment can only be obtained from the survival function estimated by AKME from the IPCW. On the other hand, when no interaction between randomized and alternative treatments exists, interpretation of causal inferences on β_1 from the MSM and IPCW is applicable to different pseudo-populations, as illustrated by the case study in Section 6.

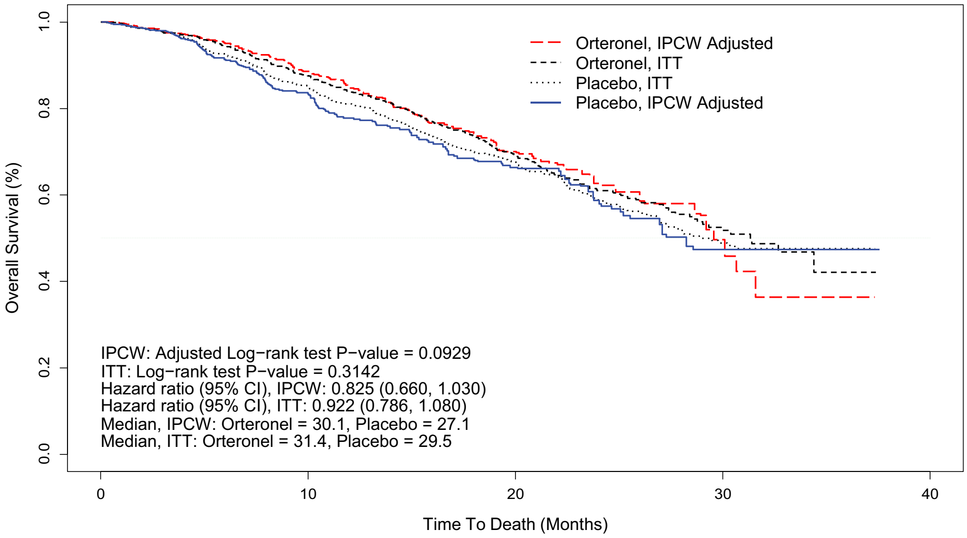

Figure 6. ELM-PC4 Trial: Overall Survival, ITT Analysis vs IPCW adjusted Analysis.

MSMs and IPCW are powerful tools that may uncover the causal effect of active treatment on OS that is otherwise obscured by the treatment switching, but powerful tools can be dangerous if not handled with care (Cole, S. R., and M. A. Hernán, 2008). Bias may be introduced, and type I error may be inflated if the measures discussed in Section 5 are not followed carefully. At Takeda, MSM and IPCW have been used to analyze confounded RCT OS data in recent years, using the following measures to reduce potential bias: pre-specification of the candidate confounder set, variable selection criteria in the weighting step, and the extreme weight truncation formula in the SAP.

Acknowledgments

We thank the referee, associate editor and editorial board for their constructive comments and suggestions, which have greatly improved this paper. The authors are also grateful to Jenny Lamont, MS, for her final review and comments of the manuscript.

Data Availability Statement

The data that support the findings of this study are available by contacting the corresponding author, upon request, and per Takeda's data sharing policy.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research received no specific grant from any funding agency

ORCID

Jing Xu http://orcid.org/0000-0002-4443-8923

References

Cole S, R., and A. Hernán M. 2004 08. Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine* 75 (1):45–49. doi:10.1016/j.cmpb.2003.10.004.

Cole, S. R., and C. E. Frangakis. 2009. The consistency statement in causal inference a definition or an assumption? *Epidemiology* 20 (1):3–5. doi:10.1097/EDE.0b013e31818ef366.

Cole, S. R., and M. A. Hernán. 2008. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 168 (6):656–664. doi:10.1093/aje/kwn164.

D'Agostino, R. B., M. L. Lee, A. J. Belanger, L. A. Cupples, K. Anderson, and W. B. Kannel. 1990. Relation of pooled logistic regression to time dependent cox regression analysis: the Framingham heart study. *Statistics in Medicine* 9 (12):1501–1515. doi:10.1002/sim.4780091214.

Farmer, R. E., D. Kounali, A. S. Walker, J. Savovic, A. Richards, M. T. May, et al. 2018. Application of causal inference methods in the analyses of randomized controlled trials: A systematic review. *MioMed Central* 19 (23):1–14. doi:10.1186/s13063-017-2381-x.

Hernán, M. Á., 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. doi:10.1097/00001648-200009000-00012.

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. doi:10.1198/016214501753168154.

Holmes, M. D., W. Y. Chen, L. Li, E. Hertzmark, D. Spiegelman, and S. E. Hankinson. 2010. Aspirin intake and survival after breast cancer. *Journal of Clinical Oncology* 28 (9):1467–1472. doi:10.1200/JCO.2009.22.7918.

Hosmer Jr. D.W., S. Lemeshow, and S. May. 2008. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. 2nd ed. Hoboken, New Jersey: John Wiley & Son's Inc., ISBN:13:978-0471754992.

Kalbfleisch, J., and R. Prentice. 2011. *The statistical analysis of failure time data*. 2nd ed. Hoboken, New Jersey: published by John Wiley & Son's Inc., ISBN:13: 978-0471363576.

Latimer, N. R., K. R. Abrams, P. C. Lambert, J. P. Morden, and M. J. Crowther. 2018. Assessing methods for dealing with treatment switching in clinical trials: A follow-up simulation study. *Statistical Methods in Medical Research* 27 (3):765–784. doi:10.1177/0962280216642264.

Rimawi, M., and S. G. Hilsenbeck. 2012. Making sense of clinical trial data: Is inverse probability of censoring weighted analysis the answer to crossover bias. *J Clin Oncol* 30 (4):453–458. doi:10.1200/JCO.2010.34.2808.

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*, 95–133. New York: Springer. ISBN-13: 978-1461270782

Robins, J. M., and D. M. Finkelstein. 2000. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 56 (3):779–788. doi:10.1111/j.0006-341X.2000.00779.x.

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

Rotnitzky, A., and J. Robins. 2005. Inverse probability weighted estimation in survival analysis. *Encyclopedia of Biostatistics* 4:2619–2625.

Saad, F., K. Fizazi, V. Jinga, E. Efstathiou, P. C. Fong, L. L. Hart, R. Jones, R. McDermott, M. Wirth, K. Suzuki, et al. 2015. Orteronel plus prednisone in patients with chemotherapy- naive metastatic castration-resistant prostate cancer (ELM-PC 4): A double-blind, multicenter, phase 3, randomized, placebo-controlled trial. *The Lancet Oncology*. 16 (3):338–348. doi:10.1016/S1470-2045(15)70027-6.

SAS Institute Inc. 2011. *SAS 9.1 User's Guide.* Cary, NC: SAS Institute Inc., ISBN:1-59047-244-6. http://www.sas.com/

Watkins, C., X. Huang, N. Latimer, Y. Tang, and E. Wright. 2013. Adjusting overall survival for treatment switches: Commonly used methods and practical application. *Pharmaceutical Statistics* 12 (6):348–357. doi:10.1002/pst.1602.

Xie, J., and C. Liu. 2005. Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine* 24 (20):3089–3110. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2174.

Yamaguchi, T., and Y. Ohashi. 2004a. Adjusting for differential proportions of second-line treatment in cancer clinical trials. part i: structural nested models and marginal structural models to test and estimate treatment arm effects. *Statistics in Medicine* 23 (13):1991–2003. doi:10.1002/sim.1816.

Yamaguchi, T., and Y. Ohashi. 2004b. Adjusting for differential proportions of second-line treatment in cancer clinical trials. part ii: an application in a clinical trial of unresectable non-small-cell lung cancer. *Statistics in Medicine* 23 (13):2005–2022. doi:10.1002/sim.1817.

Young, J. G., and E. J. Tchetgen Tchetgen. 2014. Simulation from a known cox MSM using standard parametric models for the g-formula. *Statistics in Medicine* 33 (6):1001–1014. doi:10.1002/sim.5994.