



Approaches to Selecting “Time Zero” in External Control Arms with Multiple Potential Entry Points: A Simulation Study of 8 Approaches

Anthony J. Hatswell , Kevin Deighton, Julia Thornton Snider, M. Alan Brookhart, Imi Faghmous, and Anik R. Patel

Background. When including data from an external control arm to estimate comparative effectiveness, there is a methodological choice of when to set “time zero,” the point at which a patient would be eligible/enrolled in a contemporary study. Where patients receive multiple lines of eligible therapy and thus alternative points could be selected, this issue is complex. **Methods.** A simulation study was conducted in which patients received multiple prior lines of therapy before entering either cohort. The results from the control and intervention data sets are compared using 8 methods for selecting time zero. The base-case comparison was set up to be biased against the intervention (which is generally received later), with methods compared in their ability to estimate the true intervention effectiveness. We further investigate the impact of key study attributes (such as sample size) and degree of overlap in time-varying covariates (such as prior lines of therapy) on study results. **Results.** Of the 8 methods, 5 (all lines, random line, systematically selecting groups based on mean absolute error, root mean square error, or propensity scores) showed good performance in accounting for differences between the line at which patients were included. The first eligible line can be statistically inefficient in some situations. All lines (with censoring) cannot be used for survival outcomes. The last eligible line cannot be recommended. **Conclusions.** Multiple methods are available for selecting the most appropriate time zero from an external control arm. Based on the simulation, we demonstrate that some methods frequently perform poorly, with several viable methods remaining. In selecting between the viable methods, analysts should consider the context of their analysis and justify the approach selected.

Highlights

- There are multiple methods available from which an analyst may select “time zero” in an external control cohort.
- This simulation study demonstrates that some methods perform poorly but most are viable options, depending on context and the degree of overlap in time zero across cohorts.
- Careful thought and clear justification should be used when selecting the strategy for a study.

Keywords

anchor date, big data, index date, real world data, target trial, time zero

Date received: September 28, 2021; accepted: April 4, 2022

Corresponding Author

Anthony J. Hatswell, Delta Hat, Bramley House, Bramley Road, Nottingham, NG10 3SX, UK; (ahatswell@deltahat.com).

Introduction

The use of external cohorts in regulatory or health technology assessment submissions is becoming increasingly common in the United States and Europe. A recent systematic review identified 43 occasions in which nonrandomized study designs using external controls were included in applications to US or EU regulatory authorities, most of which were met with approval.¹ Data from external comparator groups can come from a variety of sources and be used to augment clinical trials, particularly uncontrolled studies (which are often single-arm trials) when a comparison group is not feasible.

When data are taken from an external source for comparison with a clinical trial, the aim is to replicate the conditions of a randomized study as closely as possible.² Systematically attempting to do so is known as the “target trial” approach³ and involves selecting patients at the point where treatment decisions are being made, including enrolling them in the trial, should they have been available.⁴ The classical example of this design is the comparative new user design,⁵ which emulates a parallel-group randomized trial. A complexity exists, however, in which patients could be included at various points, that is, they would have been eligible for entry to the trial at several discrete time points (as opposed to only a single point). This is known in the literature as defining “time zero,” the “index date,” or “anchor date” for patients^{1,3,6,7}; here, we suggest the term “time zero” as the most appropriate and widely used term. An example of the problem is given in Hernán and Robins³ with the case of women older than 50 years of age receiving hormone therapy, who would be eligible at age 50, 51, 52, and so forth.

This issue is of particular importance in situations in which outcome risk changes depending on which potential time zero is selected; for example, in cancer patients,

prognosis generally deteriorates by treatment line.^{8–12} Consequently, any imbalance in the number of prior lines of therapy in a cross-trial comparison is likely to induce bias. Selection of eligible intervals should also be done to avoid immortal person time¹³ and other kinds of selection bias,¹⁴ where enrollment and assignment to different treatment groups depend on events that occur after the start of follow-up. For example, immortal person time occurs when patients have to survive some initial period of follow-up before they can be classified as being treated.^{13,15} In oncology, the absence of later lines of therapy is likely indicative of poor outcomes (e.g., death), with outcomes (such as response rates) likely correlated within patients.⁷

Selecting an appropriate time zero is an issue we recently faced when determining the comparative effectiveness of the chimeric antigen receptor T-cell (CAR T) therapy axicabtagene ciloleucel (Yescarta, Kite, a Gilead Company) in follicular lymphoma (FL)¹⁶ and was the motivation for completing this study. FL is an incurable disease that can recur many times within a patient’s lifetime,¹⁷ with the prognosis worsening with each passing line of therapy.¹² For the pivotal single-arm ZUMA-5 study, the experimental nature of the intervention meant that patients had to have received a minimum of 2 prior lines of therapy (the final sample had a mean of 3.6 and maximum of 9 prior lines¹⁶). To estimate comparative outcomes, an external control arm was constructed based on electronic medical records pooled with historical clinical trial data. Because of the length of time over which patients could have been included in this study, a higher proportion of earlier lines were observed, with patients generally having multiple candidate time zeros, at which point they would have fulfilled the entry criteria for the axicabtagene ciloleucel study, with an example shown in Figure 1.

In this article, we present a simulation study to add to the existing conceptual discussions for defining time zero. First, we lay out the design of the simulation, which is intended to capture the salient characteristics of the occasions when external controls are frequently used, namely, uncontrolled studies in oncology.¹⁸ We then present the various approaches available to the analyst for selecting time zero, including established and novel methods, before presenting the results of the study comparing the performance of the different methods under a range of scenarios when coupled with methods to account for confounding (i.e., propensity scoring).

Methods

Simulation Study Design

The setup of the simulation study was designed to mimic an external control following patients through multiple

Delta Hat, Nottingham, UK (AJH, KD); Department of Statistical Science, UCL, London, UK (AJH); Kite Pharma Inc, Santa Monica, CA, USA (JTS, IF, ARP); Novartis, Durham, NC, USA (MAB). The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The simulation was conducted by AJH with assistance from KD and input from all authors. The first draft of the manuscript was written by AJH and revised critically for content by all authors. All authors reviewed and approved the final version of the manuscript. The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Financial support for this study was provided by Kite, A Gilead Company and Delta Hat. JTS, IF, and ARP are employees of Kite and hold equity in Gilead. AJH and KD are employees of Delta Hat. MAB receives consulting fees and owns equity in Target RWE. He has served on scientific advisory committees for Amgen, AbbVie, Atara Biosciences, Brigham and Women’s Hospital, Kite (Gilead), NIDDK, and Vertex.

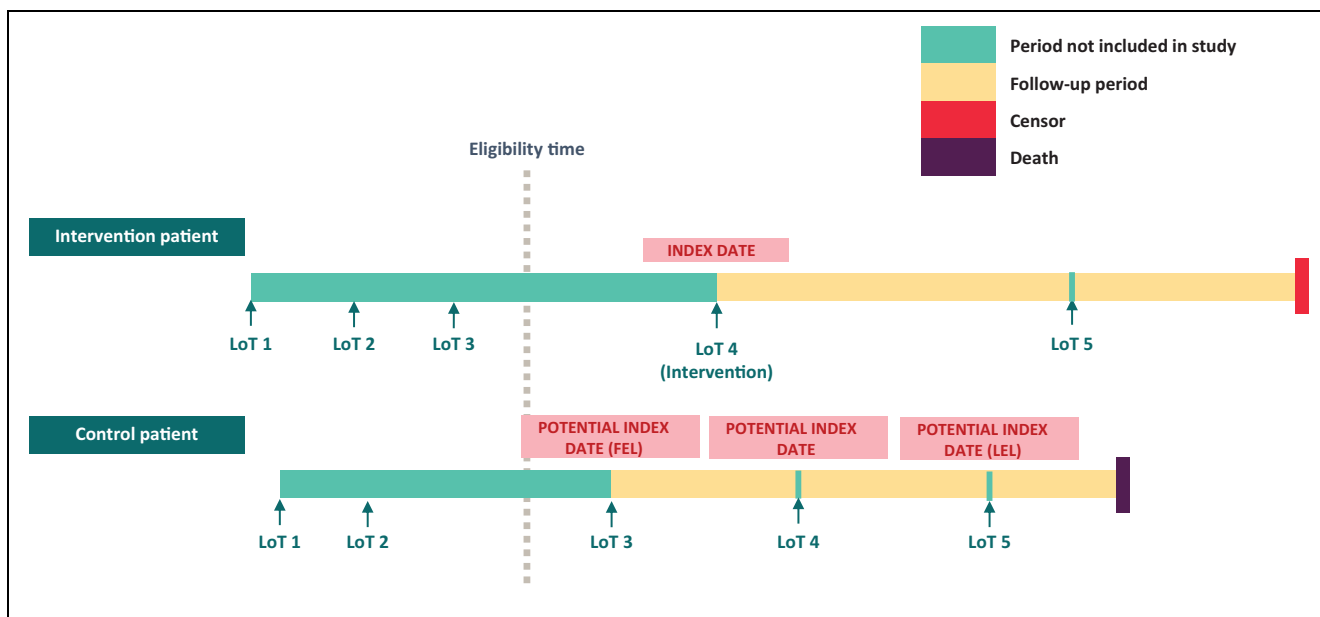


Figure 1 Stylized diagram of line selection options.

Abbreviations: LoT, Line of Therapy; FEL, First Eligible Line; LEL, Last Eligible Line.

lines of therapy and an intervention study conducted at later lines of therapy on average. Patient characteristics are sampled at the incidence of a patient's disease (i.e., line 1) and deteriorate with each line of therapy received, with each successive line also assumed to have reduced effectiveness *ceteris paribus*, as seen in many cancers.

To implement the study design, patients were deemed to have 8 characteristics (6 observable characteristics and 2 unobservable characteristics) affecting outcomes, each sampled from (independent) normal distributions, with the resulting value used to sample time-to-event outcomes from exponential distributions for time to progression and overall survival (OS), the shorter of these times then being used for progression-free survival (PFS). If a patient's first outcome was death, this time was recorded, whereas if it was progression, they moved to the next treatment line with a further set of time-to-event outcomes sampled. Patient characteristics were set to, on average, deteriorate each treatment line according to sampling from normal distributions (Figure 1). If a treatment line was deemed to be the intervention, exponential distributions with longer time-to-event outcomes were used for that line, after which they would revert to having outcomes sampled as in the external control arm. Outcomes were also set to worsen, *ceteris paribus*, as the number of prior lines a patient had received increased (Figure 2).

Starting treatment lines were drawn from a binomial distribution with 6 trials and a probability of 1/3 in the external control and 2/3 in the intervention arm (and thus

a resulting mean difference of 2 lines). This difference in starting line leads to a bias against the intervention as patients will begin treatment later in the pathway. Implicitly, this means that in a naïve comparison intervention, patients will have less favorable characteristics and thus have a worse prognosis. Within the data set, patients are then assumed to be followed up for 60 mo (external control) or 37 mo (intervention) before administrative censoring occurs. All censoring is assumed to occur at the same point so as not to introduce randomness into estimates of restricted mean survival time (RMST), which was estimated at 36 mo. A diagrammatical representation of the study is provided in Figure 1, with inputs presented mathematically in Table 1.

In the simulation, 3 data sets were constructed: the external control data set, the intervention data set, and a "true control" data set. The true control data set was a facsimile of the intervention data set to the point patients received the intervention, at which point they instead receive the control outcomes. This allows the calculation of outcomes in a set of identical patients (i.e., a true counterfactual of the same patients). Methods are then applied with the aim of estimating the effect of the intervention by comparing outcomes observed in the external control and intervention samples.

Methods for Comparison

In the simulation, 8 methods for setting time zero were implemented and used to estimate the intervention

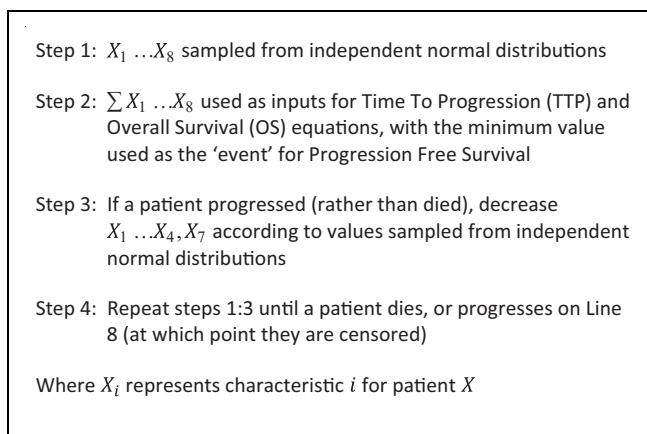


Figure 2 Diagrammatical representation of the data-generation process for patient outcomes.

effectiveness in each run. This estimated value was then compared with the results when using the values derived from the true control (i.e., those observed when calculating the effect size in the group of identical patients). Each of the methods investigated in the study for defining time zero in the external control arm are outlined below.

First eligible line. In accordance with guidance from various sources, this approach includes patient records at the first new line of therapy after meeting the eligibility criteria for the intervention study. For a study using retrospective electronic medical record data with passive qualification of patients into the study, this will generally lead to an overrepresentation of earlier lines of treatment, whereas for a prospective study, this may not be the case because of more intentional decisions on patient qualification for the study, including number of treatments failed. In the case of limited overlap in the number of prior lines of therapy between groups, we anticipate this approach to be statistically inefficient, although it should be noted that it has been used previously.^{19,20}

Last eligible line. This approach includes patient records at their last eligible line of recorded therapy. This approach has been used in empirical work,²¹ but concerns have been raised about the selection bias that it may induce in the external control group, since records are selected with the knowledge that they are the last lines of therapy received by patients and therefore more likely to end with a poor outcome.⁷

Inclusion of all lines (cloning patients who have multiple lines of treatment). This approach is discussed by

Hernán and Robins³ and involves setting the unit of analysis to individual lines of treatment, rather than individual patients, resulting in patients with multiple lines being included multiple times in the analysis. Although this approach is likely to increase statistical power, a group-based robust variance estimator must be used to estimate standard errors to address within-patient correlation of outcomes (for example, response rates).

Inclusion of all lines but censoring survival after progression. This approach is a modification of the above approach, whereby OS is censored at the point of progression between treatment lines. This modification changes the estimand but avoids having deaths attributable to multiple treatment lines observed on the same patient, an issue that superficially appears problematic.

Use of a random line of treatment. This approach is suggested by Hernán and Robins,³ which involves randomly selecting 1 line per patient in situations in which multiple eligible lines are available. The target population will reflect patients being treated in later lines than a “first eligible line” approach. This approach was discussed at length by Backenroth²² with published examples also available.^{23,24}

Rebalancing the external control arm to minimize the mean absolute error in the number of prior lines of therapy between data sets. This is a modification of the random line approach, whereby 1 line is selected per patient but with the objective of minimizing the difference in the number of prior lines between the external control and intervention data sets. This was implemented by taking 30 samples of random lines, calculating the mean absolute error in the percentage of patients at each line between the intervention and sampled external control lines, then choosing the data set with the lowest value.

Rebalancing the external control arm to minimize the root mean squared error in the number of prior lines of therapy between data sets. This is a modification of the above approach using root mean squared error (RMSE), which penalizes large mismatches in line distribution more harshly.

Using propensity score matching to identify the best matching line for each treated patient (allowing patients to be matched only once). Propensity scores are widely used in medicine to control confounding. Here, the propensity score is the probability that a given patient would be in

Table 1 Parameters Used for the Implementation of the Simulation Study

Parameter	Base-Case Value
Number of patients sampled	External control: 500 Intervention: 750
Starting line of therapy	Control: <i>Binomial</i> (probability = $\frac{1}{3}$, size = 6) Intervention: <i>Binomial</i> (probability = $\frac{2}{3}$, size = 6)
Patient characteristics ($n = 8$) at line 1	Both arms: <i>Truncatednormal</i> ($n = 8$, mean = 140, s.d. = 20, lower = 100)
Change in patient characteristics each line	Characteristics 1–3: <i>normal</i> (mean = -6, s.d. = 5,) Characteristics 4–8: <i>normal</i> (mean = 0, s.d. = 5,)
Deterioration applied by line	Applied to characteristic 8 (unobserved): <i>normal</i> (mean = -9, s.d. = 5,)
Time to progression	Control: <i>Exponential</i> $\left(3 + \frac{1}{4} \sum_{i=1}^8 x_i\right)$ Intervention: <i>Exponential</i> $\left(3 + \frac{1}{2} \sum_{i=1}^8 x_i\right)$
Overall survival	Where x_i is the vector of i characteristics for patient j Control: <i>Exponential</i> $\left(3 + \frac{1}{3} \sum_{i=1}^8 x_i\right)$ Intervention: <i>Exponential</i> $\left(3 + \frac{3}{4} \sum_{i=1}^8 x_i\right)$
Administrative censoring	Where x_i is the vector of i characteristics for patient j External control: 60 mo Intervention: 37 mo

s.d., standard deviation.

the trial versus the external control population. The propensity score is used to match eligible lines from the external control data to the nearest treated patient line using custom R code. The custom code uses a loop to select the nearest score (from all patients and lines) to a randomly selected treatment patient. After each match, that control patient (and all of their unmatched lines) are removed from the matchable pool, meaning each treated patient received a 1:1 match, but no control patients were matched more than once.

Each method was applied with and without the application of standardized mortality ratio (SMR) propensity score weighting.²⁵ This was calculated using a propensity score estimated from the 6 observable patient characteristics and the line of therapy as covariates. SMR weights were then applied whereby treated patients were given a weight of 1, whereas weights for external control patients were defined as the ratio of the estimated propensity score to 1 minus the estimated propensity score.²⁶ By using SMR weighting, we explicitly targeted the effectiveness of the intervention in the population represented in the intervention study. This provided a common reference population for all analyses. The provision of both unadjusted and adjusted results, however, does allow for understanding whether the results of an accurate comparison are due to the method used for setting time zero.

Scenario Analyses

To ensure that the results of the study are generalizable, a large number of scenario analyses was conducted, which included varying the patient numbers, simulation setup, outcomes observed, and the effect of subsequent lines of therapy. The changes made in each scenario analysis are presented in Table 2.

Outcomes Presented

The aim of the methods used is to retrieve the true intervention effectiveness when presented with an external control data set that has treatment lines biased toward earlier lines of therapy, with correspondingly better patient characteristics and outcomes. To do this, multiple estimates of effectiveness were calculated including the ratio of RMST at 3 y and a hazard ratio (HR) estimated from a Cox model.

The RMST is particularly useful in the presence of nonproportional hazards,²⁷ whereas the Cox model is frequently used in clinical studies. For each outcome, the RMSE is presented, along with the bias, for which the Monte Carlo standard error (MCSE) is also presented as a measure of variance within the results. In addition, 2 further metrics are presented for the Cox model: the coverage probability (the percentage of scenarios that contain the true value) and the error at the 95th percentile

Table 2 Scenario Analyses and Resulting Findings

Number	Scenario Setup	Findings
1	Number of patients sampled doubled in both arms	Results are consistent with the base case, although errors and coverage probabilities improve for all viable methods
2	Number of patients sampled halved in both arms	Errors are generally increased; however, no method appears disproportionately affected
3	Number of active patients halved	Errors are generally increased; however, no method appears disproportionately affected
4	Administrative censoring time halved to 18 mo	No meaningful changes in results
5	Starting health of patients increased in both arms (+12.5% at baseline)	No meaningful changes in results
6	Starting health of patients increased in the intervention only; +12.5% at baseline	Due to the (biased) comparison, naïve results are generally worse; however, post-SMR results are similar to the base case
7	More effective intervention; sampled times multiplied by 1.25	Slight improvements in coverage probabilities
8	More effective control; sum of patient characteristics multiplied by 1.25 before sampling	Slight increases in errors
9	Longer OS for both control and intervention, i.e., effect of OS reduced; sum of patient characteristics divided by 0.75 before sampling, i.e., a condition in which death is less common	OS results more uncertain for all comparisons
10	Different survival model for all time-to-event outcomes; Weibull with shape 1.25	No meaningful changes, estimates generally slightly worse
11	Different survival model for the intervention time to event estimates; Weibull with shape 1.25	Slight improvements in estimation of PFS, worsening of OS, likely driven by fewer observed events
12	Disease with a low death rate simulated; risk of the control set to that of the intervention, with sampled overall survival time multiplied by 10	PFS estimates improved for all estimates
13	Effect of health loss by treatment line doubled	Naïve estimates more inaccurate, with no meaningful changes after SMR weighting
14	Only 2 potential lines of treatment	Errors reduced, first eligible line in particular benefitting
15	Bigger imbalance between starting treatment lines; probability increased from 2/3 to 9/10 for intervention	Relative worsening of naïve errors, as well as post-SMR weighting errors for first eligible line. Relative improvements for random and rebalance approaches
16	No imbalance in treatment lines; control rate probability set equal to that of the intervention (2/3)	Relative improvement of naïve comparisons, and first eligible line; more uncertain estimates of overall survival differences
17	Intervention has no effect; all time to events set equal to that of control	All viable methods demonstrating low levels of error, with the bias present in last eligible line and all lines with censoring clear
18	Unbiased comparison; all starting lines and effectiveness calculations for the intervention set equal to the control	All viable methods demonstrating low levels of error, with the bias present in last eligible line particularly apparent

OS, overall survival; PFS, progression-free survival; SMR, standardized mortality ratio.

(as a measure of the likely maximum error). All results are presented for both PFS and OS, before, and after SMR weighting is applied.

Implementation and Software

To understand the performance of each method, a large ($n = 50,000$) number of patients were simulated for each data set (external control, intervention, and then the true

control) to understand the “true” results against which simulations would be judged. In each run, a sample of patients were then taken from the external control and intervention data sets (1000 and 750 in the base case) to which all methods of selecting time zero were applied. This process was repeated 5000 times per scenario, in line with the approach of Morris et al.²⁸

All analyses were performed in R version 4.1.2.²⁹

Results

Simulation and Base-Case Results

In the simulation, external control patients entered the study at an earlier line of therapy than intervention patients did, which created an inherent bias in patient characteristics in favor of the control, thereby reducing the observed benefit of the intervention. This is shown visually in the panels included in Figure 3 for PFS and OS across all methods for run 5000 of the base case, with base-case results provided in Table 3. Looking at PFS outcomes without the application of SMR weighting demonstrates substantial bias for all methods except for the use of propensity scoring to match similar patients. This underlines that statistical methods to account for confounding are required, regardless of the approach taken toward defining time zero.

Although the application of SMR weighting improved estimates, with most demonstrating good performance, it is immediately apparent that 2 methods, last eligible line and all lines (censoring) are biased: last eligible line in all outcomes and all lines (censoring) in OS outcomes. This finding was consistent in all simulations and scenario analyses, and thus renders these methods as nonviable. Given their poor performance and bias, the results of these 2 methods in simulations are not discussed further.

In terms of the remaining methods, the larger bias in the unweighted first eligible line approach was largely ameliorated by the application of SMR weighting. Although there were differences in the point estimates of the mean error and bias that exceeded that of the MCSE (i.e., differences beyond that seen in variability between samples), these differences did not appear meaningful between methods, given the simulated nature of the data. For example, when estimating PFS with the application of SMR weights, the bias in the Cox HR ranged from 0.005 to 0.013 across all methods, with the coverage probability of being 94.1% to 96.0%.

Although these findings of similarity between methods hold for both PFS and OS, there are differences that should be discussed, namely, that the first eligible line shows the potential for higher levels of error, as shown by its 95th percentile error being the highest of all viable methods for both PFS and OS. The other finding worth noting would be that the coverage probability for OS is notably lower for most methods. This is likely as a result of fewer observed events but should be considered.

Scenario Analysis Results

Scenario analysis setups and main findings are presented in Table 2, with distributions of the error in the ratio of

RMST for OS presented in Figure 4. Full tabulated results are available as supplementary material. The results show that as patient numbers are varied in scenario analyses (scenarios 1–3), similar results are seen to those of the base case. As would be expected, simulations with increased patient numbers perform better but without any method deviating from this pattern. This is similar with a shorter follow-up time (scenario 4), where differences are seen due to less data being available, but no method appeared better (or worse) under such circumstances.

Changing the setup of the simulation study regarding patient characteristics (scenarios 5 and 6) and the relationship between characteristics and outcomes (scenarios 7–11) again resulted in differences in the magnitude of findings, without affecting the findings themselves. Notably, all viable methods were able to account for bias in observed patient characteristics (scenario 6) after the application of SMR weighting. Findings were also not dependent on the survival modeling approach, with Weibull models used in scenarios 10 and 11 not affecting the results.

Changing the simulation to accommodate different types of disease (scenarios 12–14) resulted in only minor changes to results. Scenario 12 was designed to mimic a disease having little/no mortality impact allowing all methods still to be used for estimation of PFS. Similarly, scenarios 13 and 14 explored the impact of treatment lines, with the main finding being that with only 2 lines, first eligible line generally improved, whereas with a large gap between overlapping treatment lines, first eligible line performed poorly, with relative gains for methods than explicitly aim to rebalance.

Structural tests of methods were performed in scenarios 16–18 and demonstrated methods to be unbiased. These tests included no difference in treatment lines in scenario 16, an inert intervention in scenario 17, and a fully unbiased scenario in scenario 18. Overall, these scenarios demonstrated that after SMR weighting, treatment effects (or the lack thereof) were correctly identified, without bias being introduced that might lead to type I errors (failing to reject the null hypothesis).

Discussion

Main Findings

This simulation study explored several approaches for selecting time zero when comparing a single-arm trial to an external control cohort with an imbalance in the number of prior therapies between data sets. The number of prior therapies could predict both prognosis and the

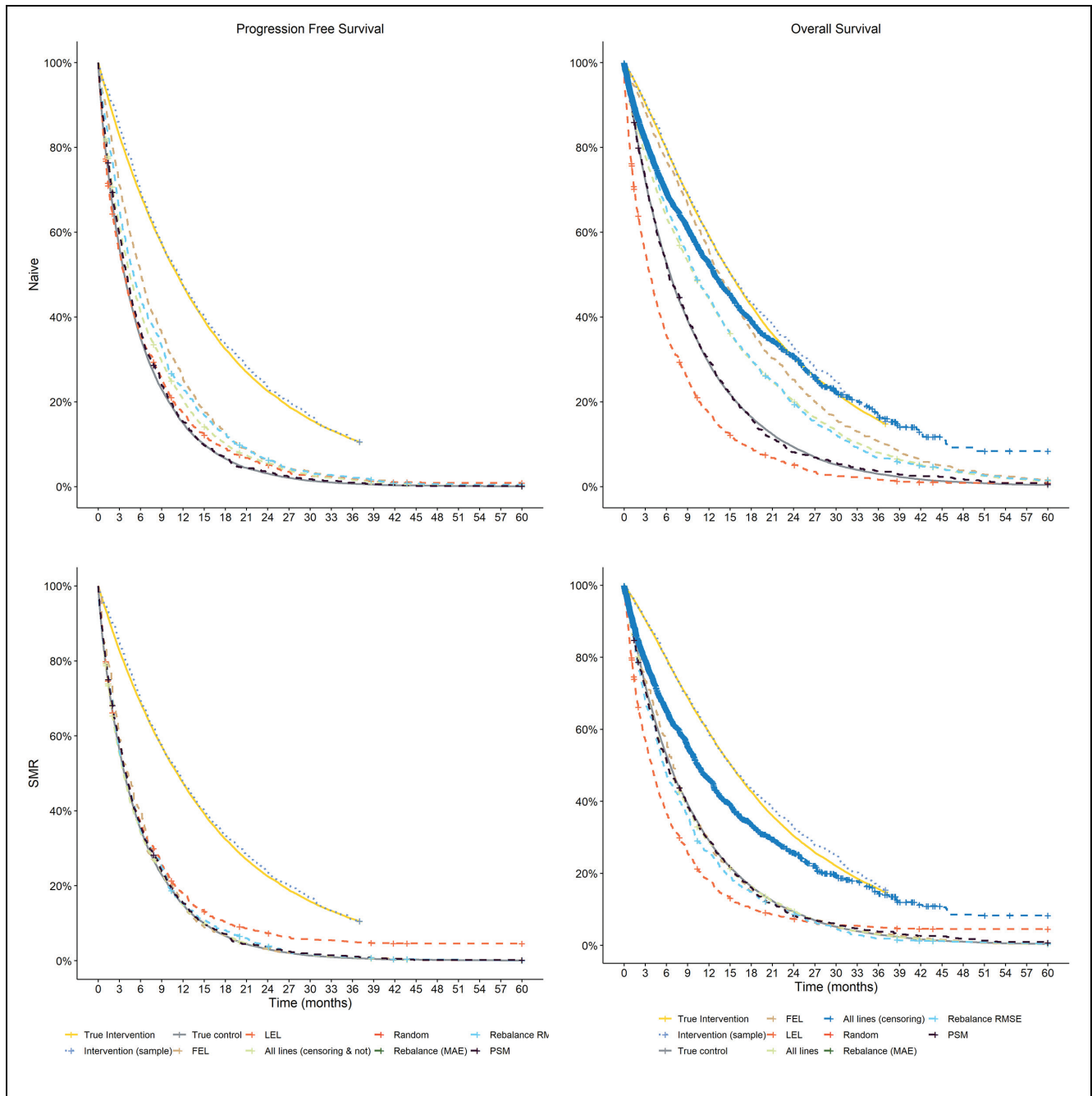


Figure 3 Example of each method applied to a single run of the simulation for progression-free survival and overall survival, with and without standardized mortality ratio weighting.

severity of covariates that deteriorate over time, which makes appropriate balance of this variable between cohorts essential when conducting comparative effectiveness studies.

Our main findings demonstrate that several methods for generating estimates in time-to-event outcomes have similar (low) levels of bias and precision, with limited evidence for a single superior approach. However, both

Table 3 Base-Case Results for Progression-Free Survival, Naïve, and Standardized Mortality Weighting Comparisons

Naïve Comparison	Progression-Free Survival												Overall Survival											
	Ratio of RMST						Cox PH Model						Ratio of RMST						Cox PH Model					
	Mean Value	RMSE	Bias (MCSE)	Mean HR	RMSE	Coverage Probability	Error 95th Percentile	Bias (MCSE)	Mean Value	RMSE	Bias (MCSE)	Coverage Probability	Mean HR	RMSE	Bias (MCSE)	Error 95th Percentile	Coverage Probability	Mean HR	RMSE	Bias (MCSE)	Error 95th Percentile	Coverage Probability		
True	2.376			0.396				1.827				0.468						0.468						
First eligible line	1.698	0.683	-0.679 (0.001)	0.54	0.147	0.144 (0)	0.191	0	1.092	0.735	-0.734 (0.001)	0.854	0.389	0.387 (0.001)	0.462	0		0.854	0.389	0.387 (0.001)	0.462	0		
Last eligible line	2.326	0.122	-0.051 (0.002)	0.42	0.032	0.024 (0)	0.06	79.8	2.836	1.017	1.009 (0.002)	0.318	0.15	-0.149 (0)	0.178	0		0.318	0.15	-0.149 (0)	0.178	0		
All lines	2.055	0.329	-0.321 (0.001)	0.469	0.075	0.073 (0)	0.106	2.5	1.347	0.482	-0.48 (0.001)	0.681	0.216	0.214 (0)	0.264	0		0.681	0.216	0.214 (0)	0.264	0		
All lines (censoring)	2.055	0.329	-0.321 (0.001)	0.469	0.075	0.073 (0)	0.106	2.5	1.347	0.482	-0.48 (0.001)	0.874	0.408	0.406 (0.001)	0.473	0		0.874	0.408	0.406 (0.001)	0.473	0		
Random	1.866	0.517	-0.51 (0.001)	0.502	0.109	0.106 (0)	0.15	0.5	1.324	0.505	-0.503 (0.001)	0.683	0.219	0.216 (0)	0.273	0		0.683	0.219	0.216 (0)	0.273	0		
Rebalanced MAE	1.882	0.502	-0.495 (0.001)	0.498	0.105	0.102 (0)	0.146	0.6	1.359	0.47	-0.468 (0.001)	0.663	0.198	0.195 (0)	0.25	0		0.663	0.198	0.195 (0)	0.25	0		
Rebalanced MSE	1.883	0.5	-0.493 (0.001)	0.498	0.105	0.102 (0)	0.145	0.6	1.367	0.462	-0.46 (0.001)	0.658	0.193	0.191 (0)	0.245	0		0.658	0.193	0.191 (0)	0.245	0		
Propensity scored	2.299	0.15	-0.077 (0.002)	0.41	0.029	0.014 (0)	0.057	88.4	1.796	0.096	-0.031 (0.001)	0.48	0.033	0.013 (0)	0.065	89.4		0.48	0.033	0.013 (0)	0.065	89.4		

Standardized Mortality Ratio Weighted	Progression-Free Survival												Overall Survival											
	Ratio of RMST						Cox PH Model						Ratio of RMST						Cox PH Model					
	Mean Value	RMSE	Bias (MCSE)	Mean HR	RMSE	Coverage Probability	Error 95th Percentile	Bias (MCSE)	Mean Value	RMSE	Bias (MCSE)	Coverage Probability	Mean HR	RMSE	Bias (MCSE)	Error 95th Percentile	Coverage Probability	Mean HR	RMSE	Bias (MCSE)	Error 95th Percentile	Coverage Probability		
True	2.376			0.396				1.827				0.468						0.468						
First eligible line	2.315	0.177	-0.061 (0.002)	0.409	0.031	0.013 (0)	0.061	96	1.773	0.114	-0.054 (0.001)	0.484	0.034	0.017 (0)	0.067	96.5		0.484	0.034	0.017 (0)	0.067	96.5		
Last eligible line	2.045	0.393	-0.332 (0.003)	0.49	0.111	0.094 (0.001)	0.172	77.9	2.478	0.697	0.651 (0.004)	0.376	0.104	-0.091 (0.001)	0.182	71.7		0.376	0.104	-0.091 (0.001)	0.182	71.7		
All lines	2.359	0.091	-0.017 (0.001)	0.401	0.018	0.005 (0)	0.037	94.5	1.779	0.077	-0.048 (0.001)	0.482	0.026	0.014 (0)	0.05	91.2		0.482	0.026	0.014 (0)	0.05	91.2		
All lines (censoring)	2.359	0.091	-0.017 (0.001)	0.401	0.018	0.005 (0)	0.037	94.5	1.773	0.655	-0.654 (0.001)	0.728	0.264	0.261 (0.001)	0.325	0		0.728	0.264	0.261 (0.001)	0.325	0		
Random	2.344	0.125	-0.033 (0.002)	0.405	0.024	0.009 (0)	0.048	93.9	1.948	0.145	0.121 (0.001)	0.44	0.035	-0.027 (0)	0.063	81.5		0.44	0.035	-0.027 (0)	0.063	81.5		
Rebalanced MAE	2.341	0.124	-0.035 (0.002)	0.405	0.024	0.009 (0)	0.047	94.1	1.966	0.161	0.139 (0.001)	0.436	0.038	-0.031 (0)	0.066	76.1		0.436	0.038	-0.031 (0)	0.066	76.1		
Rebalanced MSE	2.34	0.124	-0.036 (0.002)	0.405	0.024	0.009 (0)	0.047	94.2	1.966	0.161	0.14 (0.001)	0.436	0.038	-0.032 (0)	0.066	75.7		0.436	0.038	-0.032 (0)	0.066	75.7		
Propensity scored	2.356	0.119	-0.02 (0.002)	0.402	0.023	0.006 (0)	0.046	94.8	1.843	0.084	0.016 (0.001)	0.469	0.027	0.002 (0)	0.053	94.8		0.469	0.027	0.002 (0)	0.053	94.8		

HR, hazard ratio; MAE, mean absolute error; MCSE, Monte Carlo standard error; MSE, mean squared error; PH, proportional hazards; RMSE, root mean squared error; RMST, restricted mean survival time.

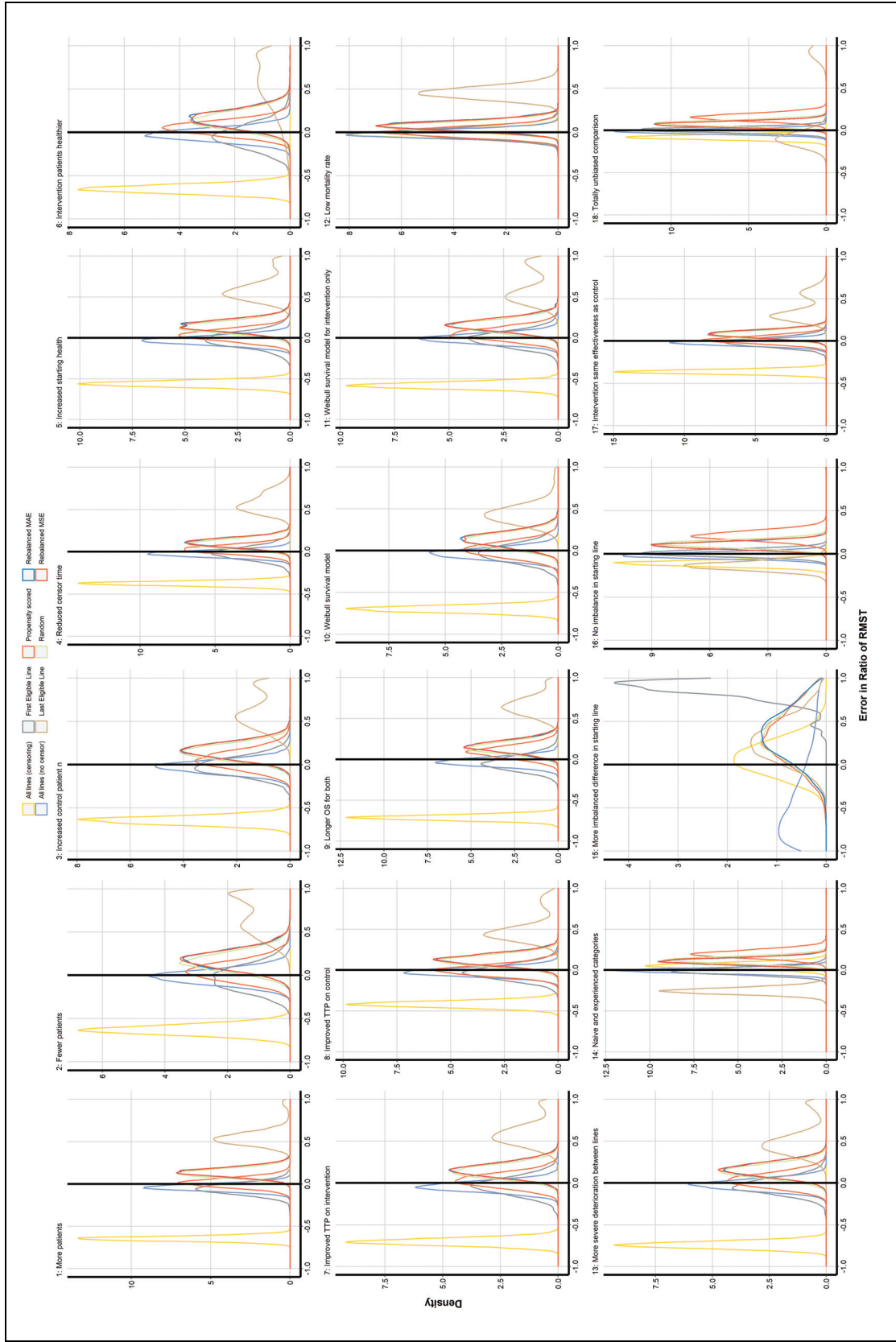


Figure 4 Density plot of error in the ratio of restricted mean survival time (RMST) for overall survival (OS), all scenario analyses.

approaches of last eligible line and the selection of all lines with censoring of OS on progression resulted in distributions of estimates that substantially deviated from our target. A key finding was that application of SMR weighting was essential in supporting the methods used for the selection of time zero, which highlights the importance of this additional consideration, namely, the use of a method to balance between groups. Following this step, multiple choices of selecting time zero could be supported, which would otherwise not be the case.

In making a selection between these methods, extensive scenario analyses indicate that the first eligible line approach may be suitable only in situations in which the external control and intervention data sets are similar or the sample sizes are large or similarly matched, due to lower statistical efficiency. Using multiple records per patient, in which death could be attributed to multiple treatment lines in the same patient, did not result in bias, and indeed frequently produced numerically superior estimates. Selecting a single random eligible line for the external control cohort also compared well with no clear bias or inaccuracy. This finding was consistent regardless of whether this process was repeated to minimize mean absolute error (MAE)/mean squared error (MSE) or performed only once. Nevertheless, repeated random sampling to minimize MAE/MSE may offer a degree of reassurance against a “bad draw” in the selection of a single random line, as can be seen in the values of 95th percentile of error. That multiple methods showed similar outcomes means that a case could be made depending on the context in which they are to be used. This includes the broader considerations of ease of explanation, perceived differences between studies, or compatibility with other methods that are also required for analysis (such as multiple imputation). The contribution of this study is therefore to present the methods that we know to be available and to identify those that should be seen as an option set from which the analyst may select. We do this based not only on theoretical advantages but also simulated data.

The bias that we found in the last eligible line approach is in agreement with the results of Suissa,⁷ who found inflated mortality rates in the control group in such a design, suggesting strong selection bias. When we censored OS on disease progression, we observed estimates that deviated substantially from our benchmark. By artificially censoring patients on progression, we essentially changed the target of estimation. In this analysis, follow-up continues only on patients who have not progressed, and they are increasingly up-weighted over time to stand in for the patients who have progressed. As such, we were implicitly estimating the effectiveness of

the intervention in a population for which progression could not occur; by artificially censoring patients on progression, we likely introduce bias from dependent censoring, because those who progress are different from those who do not. This estimator is also not of clinical relevance because progression cannot be universally prevented. An alternative approach would be to treat progression as a competing event, which would estimate the effectiveness of the intervention on death before progression. Where this issue did not occur was in scenario 12, in which survival was exogenous to the disease process, as might be seen in conditions such as migraine, psoriasis, and constipation. In this case, the censoring was not linked to outcome, with the method performing similarly to others. As we seldom fully understand disease processes, however, we would still recommend caution if suggesting this approach.

Limitations

The findings of the study were robust to extensive scenario analyses that were conducted varying parameters individually and jointly around the themes of differences in simulation setup, type of survival model used for simulation, and degree of bias (including no bias), in comparisons. A limitation, however, remains that there are further scenarios (and potentially even methods) that could be included. There are also many different decisions that could have been (legitimately) made regarding the setup of the study that may have affected the findings. The simulation has also revolved mainly around oncology products, whereas the methods available have wider applicability. Other simulation setups for different settings may therefore also be valuable.

The main limitation of the work, however, is the reliance on simulated data. Although unavoidable (in a need to have a known “truth” to compare against), this does mean that we would caution against overinterpretation of absolute results, for example, concluding one method to be superior due to lower RMSE, or bias—as it is entirely possible these small differences are an artifact of the simulation process. For these reasons, we would encourage further work, ideally in real data sets such as large randomized controlled trials from which samples could be taken and the methods compared.

Conclusions

Given the results of the simulation presented, analysts may wish to consider a variety of factors (including available sample size and degree of imbalance) in choosing an appropriate method for selecting time zero in external

cohorts. We would suggest that these include interpretability (where random line is perhaps the easiest to understand and propensity score matching or including all lines the most complex), statistical power, and interoperability with other techniques that may be required.


Beyond which method is used to set time zero, further justification should also be provided for any individual analysis, including (but not limited to) a demonstration of why each characteristic is selected for balancing, the degree of overlap between studies, histograms of weights, and effective sample sizes.

Ultimately, this study highlights a subset of methodologies with acceptable bias in the estimation of time-to-event outcomes that may be used for the selection of time zero in external control cohorts.

Acknowledgments

This work would not be possible without the freely available statistical software R and associated packages, for which the authors are extremely grateful. The authors would also like to thank Joris Diels for useful discussions regarding applications of the different approaches.

ORCID iD

Anthony J. Hatswell  <https://orcid.org/0000-0003-1129-326X>

Supplemental Material

Supplementary material for this article is available on the *Medical Decision Making* website at <http://journals.sagepub.com/home/mdm>.

References

- Mack C, Christian J, Brinkley E, et al. When context is hard to come by: external comparators and how to use them. *Ther Innov Regul Sci*. 2019;216847901987867.
- Pocock SJ. The combination of randomized and historical controls in clinical trials. *J Chronic Dis*. 1976;29:175–88.
- Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183(8):758–64.
- Brookhart MA. Counterpoint: the treatment decision design. *Am J Epidemiol*. 2015;182:840–5.
- Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol*. 2003;158:915–20.
- Wang SV, Schneeweiss S, Berger ML, et al. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1.0. *Value Health*. 2017;20:1009–22.
- Suissa S. Single-arm trials with historical controls: study designs to avoid time-related biases. *Epidemiol*. 2021;32(1):94–100. doi:10.1097/EDE.0000000000001267
- Kumar A, Sha F, Toure A, et al. Patterns of survival in patients with recurrent mantle cell lymphoma in the modern era: progressive shortening in response duration and survival after each relapse. *Blood Cancer J*. 2019;9:1–10.
- Osterlund P, Peltonen R, Alanko T, et al. A single-institution experience with bevacizumab in the treatment of metastatic colorectal cancer and in conjunction with liver resection. *OncoTargets Ther*. 2014;7:1177–84.
- Rivas-Delgado A, Magnano L, Moreno-Velazquez M, et al. Progression-free survival shortens after each relapse in patients with follicular lymphoma treated in the rituximab era. *Hematol Oncol*. 2017;35:360–1.
- Poon DMC, Wong KCW, Chan TW, et al. Survival outcomes, prostate-specific antigen response, and tolerance in first and later lines of enzalutamide treatment for metastatic castration-resistant prostate cancer: a real-world experience in Hong Kong. *Clin Genitourin Cancer*. 2018;16:402–12.e1.
- Batlevi CL, Sha F, Alperovich A, et al. Follicular lymphoma in the modern era: survival, treatment outcomes, and identification of high-risk subgroups. *Blood Cancer J*. 2020;10:74.
- Lévesque LE, Hanley JA, Kezouh A, et al. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ*. 2010;340:b5087. doi:10.1136/bmj.b5087
- Hernán MA, Sauer BC, Hernández-Díaz S, et al. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol*. 2016;79:70–5.
- Suissa S. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiol Drug Saf*. 2007;16:241–9.
- Gribben J. A comparison of clinical outcomes from ZUMA-5 (axicabtagene ciloleucel) and the international SCHOLAR-5 external control cohort in relapsed/refractory follicular lymphoma (R/R FL). 2021. Available from: https://library.ehaweb.org/eha/2021/eha2021-virtual-congress/330174/john.gribben.a.comparison.of.clinical.outcomes.from.zuma-5.28axicabtagene.html?f=menu%3D6%2Abrrowseby%3D8%2Asortby%3D2%2Amedia%3D3%2Ace_id%3D2035%2Amarker%3D1284%2Afeatured%3D17286 (accessed June 21, 2021).
- Dada R. Diagnosis and management of follicular lymphoma: a comprehensive review. *Eur J Haematol*. 2019;103:152–63.
- Goring S, Taylor A, Müller K, et al. Characteristics of non-randomised studies using comparisons with external controls submitted for regulatory approval in the USA and Europe: a systematic review. *BMJ Open*. 2019;9:e024895.
- Neelapu SS, Locke FL, Bartlett NL, et al. A comparison of one year outcomes in ZUMA-1 (axicabtagene ciloleucel) and SCHOLAR-1 in patients with refractory, aggressive non-hodgkinlymphoma (NHL). *Blood*. 2017;130:579.
- Rambaldi A, Ribera J-M, Kantarjian HM, et al. Blinatumomab compared with standard of care for the treatment of adult patients with relapsed/refractory Philadelphia chromosome-positive B-precursor acute lymphoblastic leukemia. *Cancer*. 2020;126:304–10.

21. Gökbuget N, Kelsh M, Chia V, et al. Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. *Blood Cancer J*. 2016;6:e473.
22. Backenroth D. How to choose a time zero for patients in external control arms. *Pharm Stat*. 2021;20:783–92.
23. Hansson L, Asklid A, Diels J, et al. Ibrutinib versus previous standard of care: an adjusted comparison in patients with relapsed/refractory chronic lymphocytic leukaemia. *Ann Hematol*. 2017;96:1681–91.
24. Salles G, Bachy E, Smolej L, et al. Single-agent ibrutinib in RESONATE-2™ and RESONATE™ versus treatments in the real-world PHEDRA databases for patients with chronic lymphocytic leukemia. *Ann Hematol*. 2019;98:2749–60.
25. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*. 2003;14:680–6.
26. Brookhart M, Alan, Wyss Richard, Layton J, Bradley, et al. Propensity score methods for confounding control in nonexperimental research. *Circ Cardiovasc Qual Outcomes*. 2013;6:604–11.
27. Huang B. Some statistical considerations in the clinical development of cancer immunotherapies. *Pharm Stat*. 2018;17:49–60.
28. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38:2074–102.
29. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna (Austria): R Foundation for Statistical Computing; 2020. Available from: <https://www.R-project.org>