

RESEARCH

Open Access



The impact of different censoring methods for analyzing survival using real-world data with linked mortality information: a simulation study

Wei-Chun Hsu¹, Aaron Crowley¹ and Craig S. Parzynski^{1*}

Abstract

Background Evaluating outcome reliability is critical in real-world evidence studies. Overall survival is a common outcome in these studies; however, its capture in real-world data (RWD) sources is often incomplete and supplemented with linked mortality information from external sources. Conflicting recommendations exist for censoring overall survival in real-world evidence studies. This simulation study aimed to understand the impact of different censoring methods on estimating median survival and log hazard ratios when external mortality information is partially captured.

Methods We used Monte Carlo simulation to emulate a non-randomized comparative effectiveness study of two treatments with RWD from electronic health records and linked external mortality data. We simulated the time to death, the time to last database activity, and the time to data cutoff. Death events after the last database activity were attributed to linked external mortality data and randomly set to missing to reflect the sensitivity of contemporary real-world data sources. Two censoring schemes were evaluated: (1) censoring at the last activity date and (2) censoring at the end of data availability (data cutoff) without an observed death. We assessed the performance of each method in estimating median survival and log hazard ratios using bias, coverage, variance, and rejection rate under varying amounts of incomplete mortality information and varying treatment effects, length of follow-up, and sample size.

Results When mortality information was fully captured, median survival estimates were unbiased when censoring at data cutoff and underestimated when censoring at the last activity. When linked mortality information was missing, censoring at the last activity date underestimated the median survival, while censoring at the data cutoff overestimated it. As missing linked mortality information increased, bias decreased when censoring at the last activity date and increased when censoring at data cutoff.

Conclusions Researchers should consider the completeness of linked external mortality information when choosing how to censor the analysis of overall survival using RWD. Substantial bias in median survival estimates can occur if an inappropriate censoring scheme is selected. We advocate for RWD providers to perform validation studies of their mortality data and publish their findings to inform methodological decisions better.

*Correspondence:

Craig S. Parzynski

Craig.Parzynski@genesirg.com

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Keywords Simulation, Real-world data, Real-world evidence, Survival analysis, Censoring, Missing data, Bias, Estimation

Background

The use of real-world data (RWD) to generate real-world evidence (RWE) has been increasing since it was first introduced in the 21st Century Cures Act in 2016 and, more recently, with the issuance of a series of guidances and frameworks for industry on aspects of RWD and RWE's role in regulatory decision-making [1–7]. As a result, RWE has played an increasing role in complementing randomized clinical trials (RCTs) to support new intended labeling claims, such as approval of new indications for approved medical products or in satisfying post-approval study requirements [8–15].

RWE studies typically leverage retrospective databases collected from routine practice, such as electronic health records (EHR) and administrative claims data, to answer research questions of interest. To adequately answer these questions, researchers must identify fit-for-purpose data. Several guidelines discuss the elements that should be evaluated to ensure the data chosen supports answering the research question [3, 4, 16]. Given that the data source contains relevant information vital to the research question, reliability is critical, including the data's completeness, consistency, and validity. Most importantly, researchers using RWD should ensure that outcomes are well-defined and reliably captured. The European Medicines Agency (EMA) noted inadequately captured outcomes as a top reason for the lack of study feasibility [6]. Meanwhile, the Food and Drug Administration (FDA) specifically focused on outcome misspecification, noting that its occurrence is common in RWD, and researchers should consider the potential impact of outcome misclassification on study validity [4].

Submissions using RWD/RWE to support regulatory decisions have evaluated overall survival (OS) as an important clinical outcome [8, 17, 18]. However, in contrast to prospective studies where investigators closely monitor patients, enabling the accurate collection of mortality information, the capture of survival status and timing of deaths is incomplete in RWD, which can be attributed to the data-generating process. For example, many RWD providers source data from open networks, containing only a portion of a patient's medical records. As a result, RWD sources only capture deaths occurring during the data generation process, such as disenrollment due to death from the claims database or in-hospital death captured in EHR. Consequently, RWD is typically augmented with mortality information by linking to external sources such as the

Social Security Death Index, obituaries, or the Surveillance, Epidemiology, and End Results (SEER) Program. Depending on the source and linkage mechanism, inherent properties such as data quality, data recency, and information lag can impact the reliability and accuracy of mortality data.

Further complicating data evaluation, most RWD providers do not provide information on the reliability and validity of mortality information in their documentation, highlighting a knowledge gap necessary to identify fit-for-purpose data. To our knowledge, only one data provider has explicitly published this information. In this widely used nationwide EHR-derived oncology database, mortality information ranged in sensitivity from 83.9 to 91.5% and specificity from 93.5 to 99.7% across 18 cancer types when benchmarked against the National Death Index (NDI) [19]. Without knowing the completeness and accuracy of mortality information, researchers cannot assess the amount of misspecification or its impact on the target estimand of interest, typically median survival times for each treatment and a comparison based on a hazard ratio (HR). This data gap represents an imminent risk to the validity of RWE studies using OS as an endpoint due to bias.

When mortality information is incomplete, the status of patients and actual time at risk are unknown, causing bias to enter the analysis from the misspecification of the right censoring mechanism. This occurrence is called 'ghost-time' in the literature, as the time at risk is inappropriately accrued after a patient's death due to the missing mortality information. Ghost-time can influence the median OS estimates, and the direction of this bias on relative effect measures can be upwards, downwards, or away from the null, depending on specific circumstances [20–22]. Carrigan et al. noted that estimates of median OS could be substantially biased when the sensitivity of mortality is in the 60–70% range [23]. Though the bias introduced by ghost-time is well documented, the decision of when to censor patients due to missing mortality information is less clear. In the general time-to-event (TTE) analysis framework, Lesko et al. evaluated the impact of various censoring schemes on estimating TTE endpoints with complete endpoint information. They concluded that censoring should be chosen based on the nature of the endpoint [24]. It was recommended to censor at the last study encounter for endpoints measured during a study encounter (e.g., at the last observed study

visit for the change in baseline value of a biomarker requiring a lab test result) and censor at a predefined lost to follow-up for endpoints captured outside of a study encounter (e.g., one year after the last observed study visit for death) where patients are still hypothetically at risk even if the event hasn't occurred [14]. A composite censoring scheme that adds up the cumulative hazard specific to each event type was proposed for a composite endpoint collected from within or outside the study visit [25].

While Lesko et al. illustrated bias associated with different censoring schemes, they did not consider missing information on the outcome common to RWD. We extend this previous work by performing simulations to evaluate the bias in estimates of median OS and log HRs in cases where death events are misspecified due to missing information using two different censoring schemes: (1) censoring at the last activity in the database and (2) administrative censoring at the end date of database collection (censoring at data cutoff).

Methods

Data generation

We used Monte Carlo simulation to emulate a non-randomized comparative effectiveness study of two treatments using RWD from EHR and linked external mortality data. The aims, data-generating mechanisms, estimands, methods, and performance measures (ADEMP) structural approach introduced by Morris was utilized in planning, executing, and reporting [26]. The data-generating mechanism was informed by past RWE studies in non-small cell lung cancer (NSCLC). Time to death and time to last (database) activity dates were assumed to follow an exponential distribution (Table 1). Aligned with RWD, the generated last activity date was defined as the date of a patient's last encounter in the healthcare system. Time to data cutoff, when data is no longer available, was assumed to follow a uniform distribution from 0 to 5 years. Patients were considered to have died within the data-generating process if their time to death preceded their time to the last activity date. Alternatively, patients were assumed to die outside the data-generating process when their death date was beyond their last activity date. Death information captured outside the data generation process was considered 'linked external mortality data'. No death could be observed past the data cutoff since it represented the end of any data collection. Independent censoring was assumed, where no baseline covariates informed time to the last activity date.

Estimands

We estimated the median OS and log HRs in each simulated dataset, comparing the treatment groups using Kaplan-Meier product-limit estimators and Cox proportional hazards models, respectively. The model only incorporated the treatment variable and did not consider additional covariates.

Scenarios

We varied the completeness of data capture in the linked external mortality data by randomly selecting death events (true positives) and then setting those observations to missing (false negatives). By randomly setting death observation to be missing, the linked mortality data is essentially missing completely at random (MCAR). We increased the probability of missing a death from 0 to 50%, where a probability of 0% meant that all linked mortality events were captured, and 50% indicated that only half of the linked mortality events could be captured. A probability of 0% missing death corresponded to a sensitivity of 1, while a probability of 50% missing death corresponded to a sensitivity as low as 87% across all simulation settings. The sensitivity was calculated as the proportion of correctly identified deaths relative to the total number of actual deaths (true positives and false negatives, represented by the number of deaths missing external mortality information). As a false negative corresponded to death that would have been observed had it not been missing, sensitivity decreased as the probability of missing death increased. The accuracy of death dates and specificity of linked external mortality captured were not varied within our study.

We anticipated that sample size, amount of follow-up, and treatment effect size could influence the amount of bias. Our primary analysis generated 400 patients assigned in a 1:1 ratio to the treatment and control groups with a null ($HR=1$) and moderate treatment effect ($HR=0.75$). In the primary analysis, the median OS was simulated to be 1.5 years for the control group and set to 1.5 years for the treatment group under a null treatment effect and 2.0 years for the treatment group under the moderate treatment effect. The median OS for the control was chosen based on the median OS observed among NSCLC patients from RWD, but this could also represent other cancers with worse prognosis, such as hepatocellular carcinoma, advanced head and neck cancer, or malignant pleural mesothelioma, which have median OS in the range of 1.2 to 1.6 years [19]. The treatment effect was determined based on the mean HRs for novel cancer treatments [27]. The median length of follow-up was informed by RWD among NSCLC patients and was set to 2.5 years, corresponding to 46% of patients

Table 1 Simulation setting for the primary and sensitivity analyses

Scenario	Number of patients	Time to last activity*	Time to data cutoff*	Time to death*	Number of repetitions
1. Small Sample, Short Follow-Up					
	400 patients were assigned to the treatment and control groups in a 1:1 ratio. Treatment assignment was generated using $T_i \sim \text{Bern}(0.5)$.	Follows an exponential distribution, with a median time to the last activity date being 2.5 years for both treatment groups. Time to last activity was generated using $X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right)$, where $\lambda_i = \frac{\ln(2)}{2.5}$ is the baseline hazard rate.	Follows a uniform distribution and is truncated at 5 years. Time to data cutoff was generated using $D_i \sim U(0, 5)$.	HR=1: Follows an exponential distribution. Median OS was 1.5 years for both the control and treatment groups. **Time to death was generated using	4000
				$X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right) \text{ and } \lambda_i = \begin{cases} \frac{\ln(2)}{\frac{1}{1.5}}, & T_i = 0 \\ \frac{\ln(2)}{\frac{1}{1.5}} \times 1, & T_i = 1 \end{cases}$ <p>where T_i is the treatment assignment.</p>	
				HR=0.75: Follows an exponential distribution. Median OS was 1.5 years for control and 2 years for treatment.** Time to death was generated using	6000
				$X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right) \text{ and } \lambda_i = \begin{cases} \frac{\ln(2)}{\frac{1}{1.5}}, & T_i = 0 \\ \frac{\ln(2)}{\frac{1}{1.5}} \times 0.75, & T_i = 1 \end{cases}$ <p>where T_i is the treatment assignment.</p>	
				HR=0.5: Follows an exponential distribution. Median OS was 1.5 years for control and 3 years for treatment. Time to death was generated using	5000
				$X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right) \text{ and } \lambda_i = \begin{cases} \frac{\ln(2)}{\frac{1}{1.5}}, & T_i = 0 \\ \frac{\ln(2)}{\frac{1}{1.5}} \times 0.5, & T_i = 1 \end{cases}$ <p>where T_i is the treatment assignment.</p>	

Table 1 (continued)

Scenario	Number of patients	Time to last activity*	Time to data cutoff*	Time to death*	Number of repetitions
2. Small Sample, Extended Follow-Up	400 patients were assigned to the treatment and control groups in a 1:1 ratio. Treatment assignment was generated using $T_i \sim \text{Bern}(0.5)$.	Follows an exponential distribution, with a median time to the last activity date being 6.4 years for both treatment groups. Time to last activity was generated using $X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right)$, where $\lambda_i = \frac{\ln(2)}{6.4}$ is the baseline hazard rate.	Follows a uniform distribution and is truncated at 5 years. Time to data cutoff was generated using $D_i \sim U(0, 5)$.	HR=1: Follows an exponential distribution. Median OS was 1.5 years for both the control and treatment groups. Time to death was generated using	2000
				$X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right)$ and $\lambda_i = \begin{cases} \frac{\ln(2)}{1.5}, & T_i = 0 \\ \frac{\ln(2)}{1.5} \times 1, & T_i = 1 \end{cases}$ where T_i is the treatment assignment.	5000
				HR=0.75: Follows an exponential distribution. Median OS was 1.5 years for control and 2 years for treatment. Time to death was generated using	5000
				$X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right)$ and $\lambda_i = \begin{cases} \frac{\ln(2)}{1.5}, & T_i = 0 \\ \frac{\ln(2)}{1.5} \times 0.75, & T_i = 1 \end{cases}$ where T_i is the treatment assignment.	4000
				HR=0.5: Follows an exponential distribution. Median OS was 1.5 years for control and 3 years for treatment. Time to death was generated using	4000
				$X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right)$ and $\lambda_i = \begin{cases} \frac{\ln(2)}{1.5}, & T_i = 0 \\ \frac{\ln(2)}{1.5} \times 0.5, & T_i = 1 \end{cases}$ where T_i is the treatment assignment.	

Table 1 (continued)

Scenario	Number of patients	Time to last activity*	Time to data cutoff*	Time to death*	Number of repetitions
3. Large Sample, Short Follow-Up	2000 patients were assigned to the treatment and control groups in a 1:1 ratio. Treatment assignment was generated using $T_i \sim \text{Bern}(0.5)$.	Follows an exponential distribution, with a median time to the last activity date being 2.5 years for both treatment groups. Time to last activity was generated using $X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right)$, where $\lambda_i = \frac{\ln(2)}{2.5}$ is the baseline hazard rate.	Follows a uniform distribution and is truncated at 5 years. Time to data cutoff was generated using $D_i \sim U(0, 5)$.	HR=1: Follows an exponential distribution. Median OS was 1.5 years for both the control and treatment groups. Time to death was generated using	600
				$X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right)$ and $\lambda_i = \begin{cases} \frac{\ln(2)}{1.5}, & T_i = 0 \\ \frac{\ln(2)}{1.5} \times 1, & T_i = 1 \end{cases}$ where T_i is the treatment assignment.	
				HR=0.75: Follows an exponential distribution. Median OS was 1.5 years for control and 2 years for treatment. Time to death was generated using	2000
				$X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right)$ and $\lambda_i = \begin{cases} \frac{\ln(2)}{1.5}, & T_i = 0 \\ \frac{\ln(2)}{1.5} \times 0.75, & T_i = 1 \end{cases}$ where T_i is the treatment assignment.	
				HR=0.5: Follows an exponential distribution. Median OS was 1.5 years for control and 3 years for treatment. Time to death was generated using	2000
				$X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right)$ and $\lambda_i = \begin{cases} \frac{\ln(2)}{1.5}, & T_i = 0 \\ \frac{\ln(2)}{1.5} \times 0.5, & T_i = 1 \end{cases}$ where T_i is the treatment assignment.	

Table 1 (continued)

Scenario	Number of patients	Time to last activity*	Time to data cutoff*	Time to death*	Number of repetitions
4. Large Sample, Extended Follow-Up	2000 patients were assigned to the treatment and control groups in a 1:1 ratio. Treatment assignment was generated using $T_i \sim \text{Bern}(0.5)$.	Follows an exponential distribution, with a median time to the last activity date being 6.4 years for both treatment groups. Long. Time to last activity was generated using $X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right)$, where $\lambda_i = \frac{\ln(2)}{6.4}$ is the baseline hazard rate.	Follows a uniform distribution and is truncated at 5 years. Time to data cutoff was generated using $D_i \sim U(0, 5)$.	HR=1: Follows an exponential distribution. Median OS was 1.5 years for both the control and treatment groups. Time to death was generated using	200
				$X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right)$ and $\lambda_i = \begin{cases} \frac{\ln(2)}{1.5}, & T_i = 0 \\ \frac{\ln(2)}{1.5} \times 1, & T_i = 1 \end{cases}$ where T_i is the treatment assignment.	
				HR=0.75: Follows an exponential distribution. Median OS was 1.5 years for control and 2 years for treatment. Time to death was generated using	500
				$X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right)$ and $\lambda_i = \begin{cases} \frac{\ln(2)}{1.5}, & T_i = 0 \\ \frac{\ln(2)}{1.5} \times 0.75, & T_i = 1 \end{cases}$ where T_i is the treatment assignment.	
				HR=0.5: Follows an exponential distribution. Median OS was 1.5 years for control and 3 years for treatment. Time to death was generated using	2000
				$X_i \sim \text{Exp}\left(\frac{1}{\lambda_i}\right)$ and $\lambda_i = \begin{cases} \frac{\ln(2)}{1.5}, & T_i = 0 \\ \frac{\ln(2)}{1.5} \times 0.5, & T_i = 1 \end{cases}$ where T_i is the treatment assignment.	

*The parameters are presented in the unit of years. The unit was converted to days in the execution using the sim_data function
**The marked scenarios indicate the primary analysis

lost to follow-up before the data cutoff. Three additional scenarios, representing (1) a small sample size with extended follow-up, (2) a large sample size with short follow-up, and (3) a large sample size with extended follow-up, were also evaluated. The total number of patients in each treatment group was increased to 2000 in the large sample scenarios, and the median length of follow-up was set to 6.4 years in the extended follow-up scenarios, which was chosen such that 23% were lost to follow-up before the data cutoff date, which was half of the amount of the primary analysis. We also conducted all scenarios under a strong treatment effect ($HR=0.5$). For each scenario, the number of repetitions was set to achieve a Monte Carlo standard error (MCSE) of 0.1 for the primary performance measure, the bias of median OS. The detailed settings of all scenarios, including the number of repetitions, are included in Table 1.

Censoring schemes

We evaluated two censoring schemes commonly implemented for OS in RWE studies (Fig. 1) [28]. The first scheme, referred to as “censoring at the last activity”,

censors patients at their last recorded encounter in the simulated data if no death event is observed. In this scheme, any deaths that occur before the data cutoff date are always considered an event, regardless of whether they occur before or after the last recorded encounter. The second scheme, known as “censoring at data cutoff”, censors patients at the data cutoff date in the absence of a death event. This approach is often called administrative censoring.

Each of these approaches can introduce bias in the presence of misspecified death. Specifically, the censoring at the last activity scheme may censor patients who are still at risk earlier than they should be. On the other hand, the censoring at data cutoff scheme may consider patients at risk later than they should be, leading to what is known as ghost-time bias.

Performance criteria

We calculated bias, variance, coverage, and rejection rate to evaluate the performance of the two censoring schemes. Bias was defined as the difference between the estimated and true parameters. Mean bias from

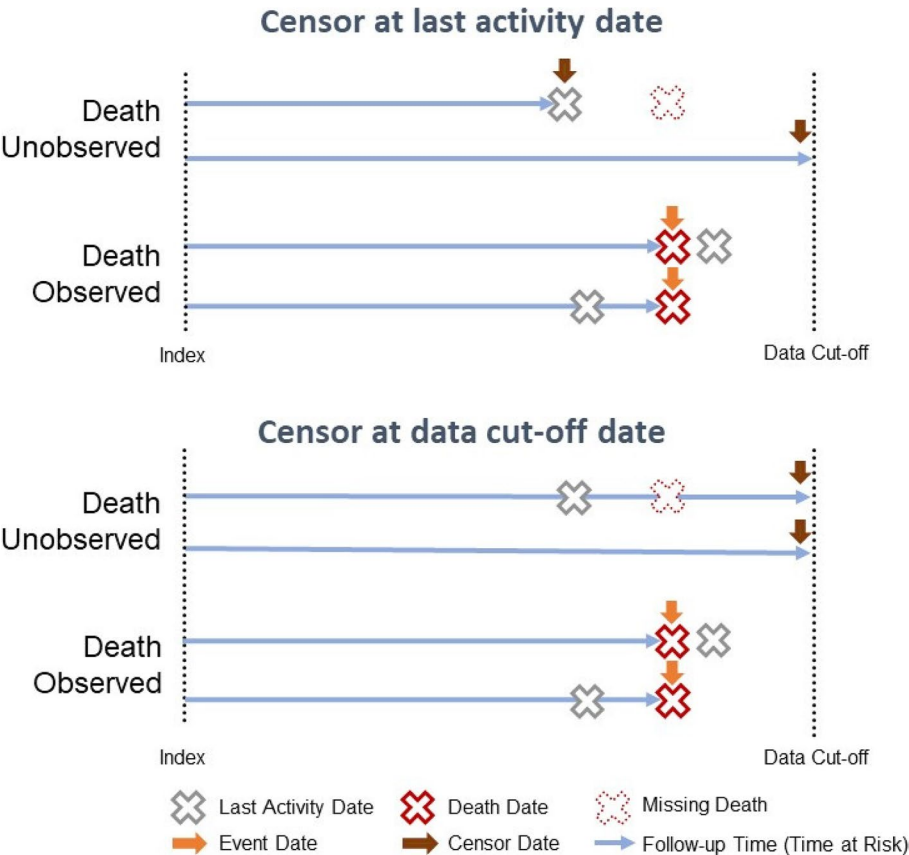


Fig. 1 Illustration of censoring schemes for analyzing OS in RWD. *The last activity date is illustrated after the death date to reflect the idiosyncrasies of RWD, where delays in data reporting and runover of clinical activities are observed and may cause discrepancies in information

the repetitions was reported. Variance was defined as the mean of squared differences from the mean of the estimated parameters. Coverage was defined as the proportion of repetitions where the estimated 95% confidence interval covered the true parameters. Rejection rate was defined as the proportion of repetitions where the null hypothesis was rejected, corresponding to the estimated type I error in the absence of treatment effect and estimated power in the presence of treatment effect. To contextualize the estimated power, expected power assuming the true treatment effect was also presented. Expected power was derived using the formula presented by Freedman and Rosner [29, 30]. MCSE of bias of median OS is included in Appendix 1, Fig. 3.

Statistical software

The study was implemented using R version 4.3.1. The simulated data was generated using a customized function, 'sim_data,' that allows for various parameter inputs related to the simulation setting (e.g., number of repetitions, median OS, HR). This function is included in Appendix 3 with a detailed explanation of the parameter settings. Performance metrics, including bias, coverage, variance, and rejection rate, were evaluated using the simhelpers package [31]. Expected power was derived

using the PowerCT function from the powerSurvEpi package.

Results

Median overall survival (OS) estimates

Estimates of median OS were unbiased when censoring at data cutoff and underestimated when censoring at the last activity in the scenario where mortality information was fully captured (Fig. 2). As expected, censoring at the last activity date led to an underestimation of the median OS, while censoring at the data cutoff resulted in an overestimation when linked mortality information was partially missing. As the probability of missing linked external mortality data increased from 0 to 50%, we observed an increase in the bias of median OS for censoring at data cutoff (mean bias in months ranged from 0.12 to 2.98 and 0.18 to 5.52 in the control and treatment groups, respectively) (Appendix 2). Conversely, the bias decreased for censoring at last activity (-1.31 to -0.72 and -2.79 to -1.59 in the control and treatment groups, respectively) (Appendix 2). The decreasing bias for censoring at the last activity can be attributed to the lower mortality subject to capture by external sources, resulting in less underestimation of time at risk for patients who would have survived until the data cutoff. Interestingly,

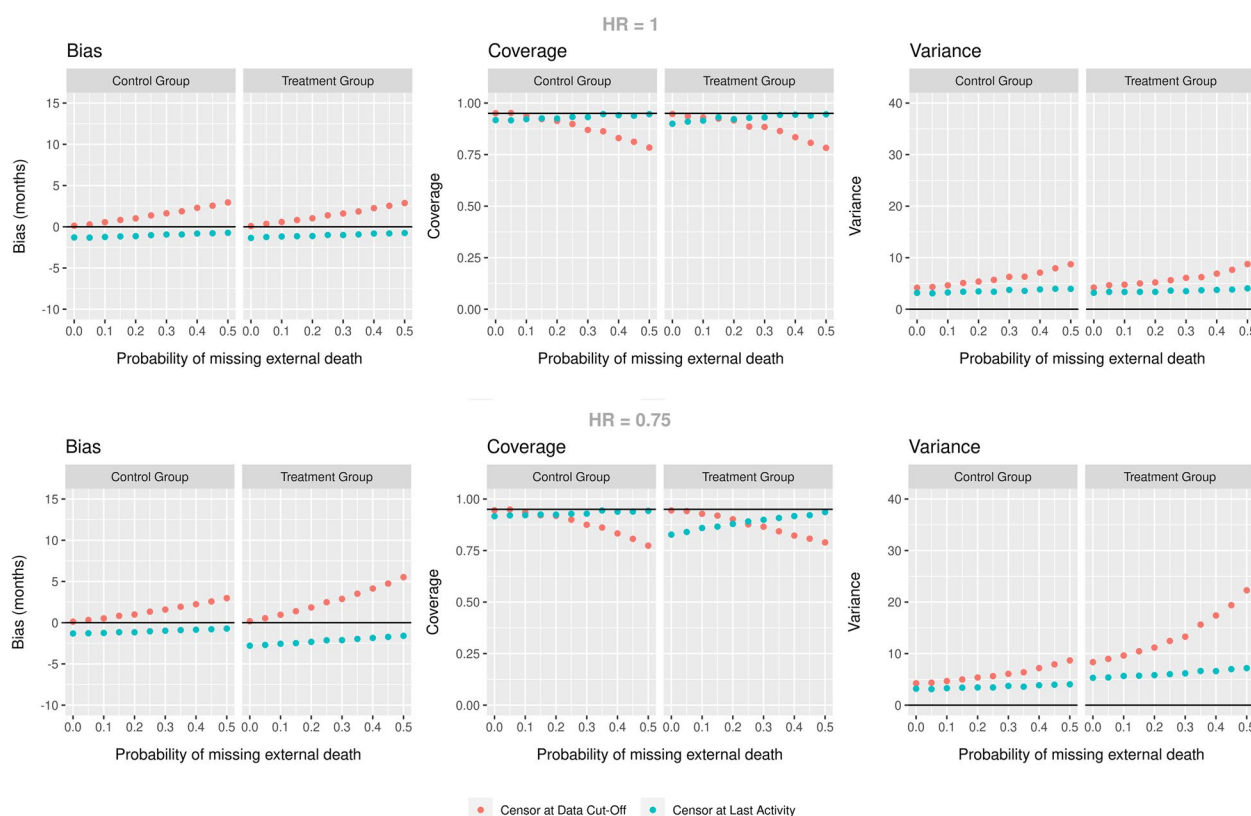


Fig. 2 Performance of median OS estimates by censoring schemes for the primary analysis

*reference line in the figure denotes 0 bias, the expected coverage or 0 variance

the magnitude of bias was smaller for censoring at data cutoff than for censoring at the last activity when the probability of missing linked external mortality data was less than 25% (Fig. 2). While the trend of bias was consistent across all scenarios, the magnitude of bias was more pronounced in the treatment group compared with the control group in the presence of treatment effect due to increased survival in the treatment group, leading to more misspecification.

Similar to bias, the coverage probabilities were maintained at 95% by censoring at data cutoff and underestimated by censoring at the last activity when mortality was fully observed. As the probability of missing linked external mortality increased from 0 to 50%, the coverage decreased away from 95% by censoring at data cutoff (mean coverage probability ranged from 95 to 75% and 94 to 79% in the control and treatment groups, respectively) and increased for censoring at last activity (92 to 94% and 83 to 94% in the control and treatment groups, respectively). In the presence of a treatment effect, a more significant deviation from 95% coverage was observed for the treatment group compared with the control group. As expected, estimates of the variance of the median OS increased as linked mortality became missing. The variance increased more for censoring at data cutoff and the treatment group when a treatment effect existed.

When the length of follow-up and sample size were varied, a shorter follow-up length increased the magnitude of bias, decreased the coverage, and increased the variance of the median OS from missing linked mortality

(Appendix 1, Fig. 1a and c). On the other hand, increasing the sample size had no impact on the bias but mitigated the effect of missing linked mortality on the coverage and variance of median OS, respectively (Appendix 1, Fig. 1a-c). These observations were also consistent for a smaller HR of 0.5, with the exception that coverage, particularly for censoring at data cutoff in the treatment group, in the short follow-up scenario was non-estimable for a majority of the repetitions, which led to uninterpretable results (See Appendix 1, Fig. 1b).

Hazard ratio (HR) estimates

Neither censoring schemes biased the estimates of log HRs nor inflated type I error in the absence of treatment effect (Fig. 3). The bias of log HRs remained close to 0, and type I error remained close to 0.05 when missing linked mortality was varied from 0 to 50%. In the presence of a treatment effect, censoring at data cutoff did not bias the estimates of log HRs when mortality was entirely captured and biased the estimates of log HRs slightly toward the null (i.e., underestimated) as missingness in linked mortality increased (mean bias in log HRs: 0% missing: -0.0004 to 50% missing: 0.02). Conversely, censoring at the last activity biased the estimates of log HRs slightly toward the null when mortality was fully captured, and the bias diminished as linked mortality became missing (mean bias in log HRs: 0% missing: 0.03 to 50% missing: 0.01) (Appendix 2). Although censoring schemes resulted in the bias of log HR, the magnitude was minimal, with a maximum bias of 0.02 for censoring at data cutoff and 0.03 for censoring at the last activity.

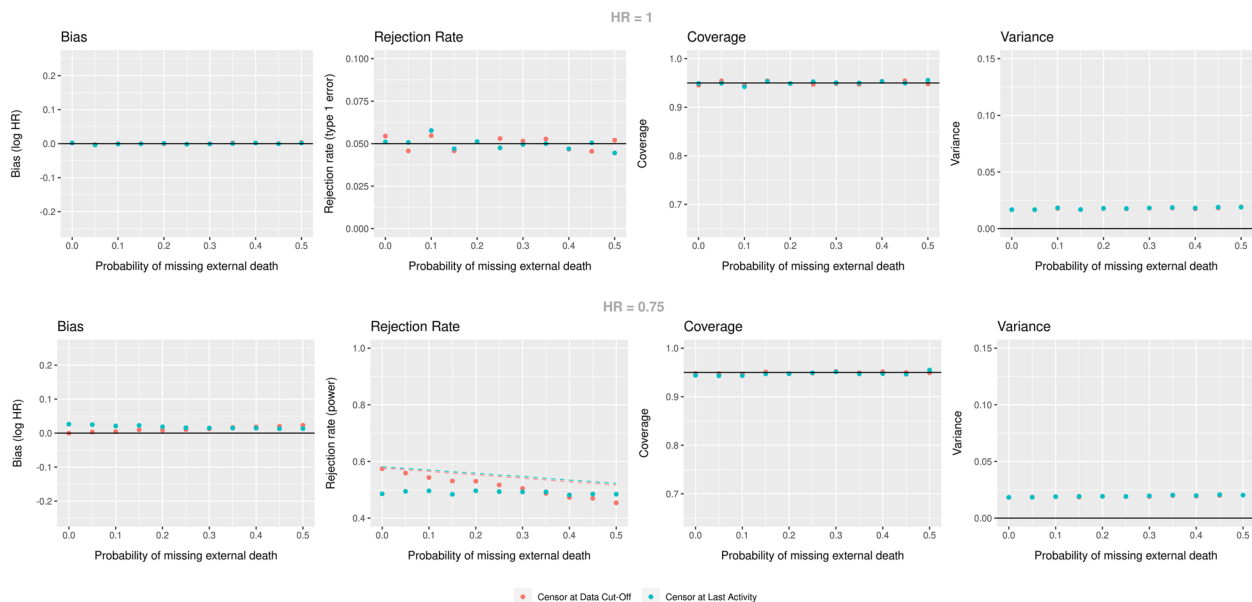


Fig. 3 Performance of log HR estimates for OS by censoring scheme for the primary analysis. *The dashed line in the rejection rate figure indicates the expected power associated with the two censoring schemes in the varying probability of missing external mortality

As the probability of missing linked external mortality data increased from 0 to 50%, the estimates of power decreased away from the expected power (mean power ranged from 57.6 to 51.6% and 58.0 to 52.2% when censoring at the data cutoff and last activity, respectively) as the estimated HRs shifted toward the null for censoring at data cutoff (mean power: 0% missing: 0.57 to 50% missing: 0.45), and the estimates of power moved toward the expected power as HRs shifted toward the true treatment effect for censoring at the last activity (mean power: 0% missing: 0.49 to 50% missing: 0.48). Interestingly, censoring at data cutoff had a minor loss of power to the expected compared to censoring at the last activity when the probability of missing linked external mortality data was less than 35%.

Estimates of coverage were maintained at the expected 95%, and variance estimates remained constant in the range of 0.016 to 0.02, irrespective of the censoring scheme, treatment effect, and probability of missing linked external mortality data.

When the length of follow-up and sample size were varied, a shorter length of follow-up increased the magnitude of bias from missing linked mortality, but not substantially. It reduced the coverage and power in the presence of treatment effect (Appendix 1, Fig. 2a-c). The variance of log HRs was not affected by the length of follow-up (Appendix 1, Fig. 2d). An increase in sample size did not affect the bias of the log HR. Still, it alleviated the effect of missing linked mortality for inference through a further reduction of variance and a minor loss in the power of HR. As coverage is inversely related to variance, the reduction in variance due to a larger sample size led to more significant variability and lower coverage of the log HR. These observations were consistent with a smaller HR of 0.5. Note that the effect of length of follow-up and sample size was merely observable with a smaller HR. Due to the enormous magnitude of the treatment effect, the estimated power was always 1, irrespective of missing linked mortality in the extended follow-up and large sample size scenario (Appendix 1, Fig. 2c).

Discussion

We used simulations to evaluate the performance of different censoring schemes under varying degrees of loss in linked mortality information. Different scenarios with varying treatment effects, lengths of follow-up, and sample sizes were considered to represent real-world data scenarios and assess how these parameters affect the choice of censoring schemes.

Our findings are consistent with previous literature. Censoring at data cutoff produces unbiased estimates for the median OS when linked mortality is fully captured. This censoring scheme accurately accounts for patients

at risk beyond their last study encounter. Conversely, censoring at the last activity underestimates the median OS as it incorrectly excludes patients still at risk. These findings correspond to the concept of a ‘captured’ outcome, as described by Lesko et al., where patients should be censored at a predefined loss to follow-up time [25]. Carrigan et al. [23] also observed an overestimation of median OS when adopting a censoring scheme similar to censoring at data cutoff. Over- or underestimation of the median OS is the result of incorrectly censoring a patient when they should be considered at risk. Patients are prematurely censored when censoring at last activity date even though they may still be at risk, or conversely contribute ghost-time to the analysis when their death is missing and censoring at data cutoff. Either approach will result in bias when mortality information is linked, but not fully captured.

Interestingly, while median OS was found to be biased due to missing linked mortality, estimates of HR remained relatively unaffected even when the sensitivity of mortality was lower in the 60–70% range. This aligns with our results, showing minimal bias in HR estimates regardless of the censoring scheme and missing linked mortality. Given that the sensitivity of mortality did not drop below 87% in our simulation, we anticipate HR to be close to the truth as the bias introduced by censoring is not dramatically different between treatment and control groups.

The relative performance of the two commonly recommended censoring schemes in analyzing RWD has yet to be formally explored. Our findings suggest that when linked mortality is moderately misspecified (missing below the 25–30% range), censoring at data cutoff is associated with a more negligible bias for median OS. Conversely, when missing above 25–30%, censoring at the last activity produces a more negligible bias for the median OS.

It's worth noting that the amount of bias is expected to decrease for censoring at the last activity as missingness in linked mortality increases. We expect an unbiased estimate when linked mortality is completely missing, which is due to the expectation that all deaths would occur within the data generation process (i.e., before the last activity date) and the time at risk would be accurately assigned for patients with and without death by censoring at the last activity date.

The length of follow-up was shown to impact the amount of bias introduced. Shorter follow-up times were associated with more bias than longer follow-up times, as early censoring led to more death misspecification from linked mortality. Censoring schemes introduce minimal bias to log HR estimates, and the magnitude is considered negligible (e.g., maximum absolute bias < 0.07).

For studies concerning statistical inference, censoring at data cutoff has a minor loss in power compared with censoring at the last activity when linked mortality is missing below the range of 30–40%. As with bias in median OS, the relative performance of the two censoring schemes flips when the probability of missing linked external mortality data is beyond this range.

The results of our study suggest that estimates of median OS obtained from RWD are sensitive to methodological decisions. It is imperative to understand the data-generation process and the accuracy and completeness of the mortality data when planning RWE studies to make appropriate decisions. Considerations should include the degree to which external mortality data is leveraged, the quality and completeness, and the limitations of the linkage mechanisms. Upon these assessments, researchers can then gauge the impact of such mortality data on study results and choose the correct methods to minimize the potential bias. The FDA recommends including a validation approach that details the design, methods, and processes for study outcomes and conducting a sensitivity analysis that evaluates the impact of outcome misspecification [4]. To adequately meet these recommendations, data providers should share validation information for mortality data and other relevant outcomes, allowing researchers to plan analyses and mitigate bias accordingly.

There are several limitations when interpreting the study findings. First, the estimates on OS were generated based on specific parameter settings informed by previous RWE studies and may not be generalizable to every study population. However, we explored the impact of a broad range of parameters (i.e., treatment effect, length of follow-up, and sample size) on key estimands in survival that should apply to many populations. In these settings, the relative performance of the two censoring schemes should persist. However, the threshold of missing linked mortality at which we determined one censoring scheme is superior may change. Our study also did not consider informative censoring. The likelihood of a patient being lost to follow-up (e.g., date of last activity) was unrelated to baseline characteristics. The bias observed in our study could be increased or decreased if patient characteristics were related to why patients are lost to follow-up in RWD. Our study did not consider issues with data recency and assumed mortality information captured from linkage was complete and up to date. An alternative data cutoff date at which mortality is considered up to date should be selected depending on the specific data source. Our study only evaluated the impact of missing linked mortality information (i.e., false negatives) and did not assess the misspecification of death dates or false indications of death events. Different results could be

observed if these were misspecified. However, reported specificity in EHR has been higher than sensitivity, indicating false negatives are more common in RWD [19]. Finally, our study assumed that the misspecification of linked external mortality data was random. This reflects an ideal scenario where the linked mortality data is complete and representative of the overall population. Still, RWD may not accurately collect linked external mortality data from a representative source, and this may not reflect actual representativeness of linked external mortality information.

Conclusion

When analyzing OS, our study suggests that commonly recommended censoring schemes may both be appropriate depending on completeness of linked mortality information captured. Censoring at data cut-off is the correct choice when linked mortality is completely captured, while censoring at the last activity date is the correct choice when no linked mortality is available. In the presence partially missing linked mortality, bias is introduced into the median OS estimate because of the hybrid collection of mortality data. We observed that censoring at data cutoff resulted in a more negligible bias for median OS and a minor loss in power for HR when the mortality is sufficiently captured, while censoring at the last activity date had less bias when mortality was not well captured. However, both censoring schemes resulted in negligible bias in log HR even when linked mortality was relatively incomplete. In real-world applications where linked mortality information is likely to be partially missing, the choice of when to censor becomes a tradeoff in properly specifying the risk sets. Understanding of how well the linked information is captured becomes critical in making the correct decision. As a result, we advocate for RWD providers to perform validation studies of their mortality data and to publish their findings so that researchers can understand the reliability of mortality data and make appropriate methodological decisions.

Abbreviations

ADEMP	aims, data-generating mechanisms, estimands, methods, and performance measures
EHR	electronic health records
EMA	European Medicines Agency
FDA	Food and Drug Administration
HR	hazard ratio
MCSE	Monte Carlo standard error
NDI	National Death Index
NSCLC	non-small cell lung cancer
OS	overall survival
RCTs	randomized clinical trials
RWD	real-world data
RWE	real-world evidence
SEER	Surveillance, Epidemiology, and End Results
TTE	time-to-event

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12874-024-02313-3>.

Supplementary Material 1.

Authors' contributions

WCS developed the study concept, conducted simulation analyses, interpreted findings and was a major contributor in writing the manuscript. AC developed the study concept, conducted simulation analyses, interpreted findings and was a major contributor in writing the manuscript. CP developed the study concept, conducted simulation analyses, interpreted findings and was a major contributor in writing the manuscript.

Funding

No external funding was received for the conduct of this study.

Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

This study involves the use of simulated data. It does not involve human participants, human data, or human tissues. Consequently, it does not require ethics board approval or consent to participate according to the guidelines of the Investigational Review Board.

The simulated data used in this study were generated to ensure they do not resemble or can be linked to any real individuals or groups. As such, the study poses no risk of harm or breach of privacy to any individual or group.

Consent for publication

Not applicable.

Competing interests

WCS, AC, and CP are employees of Genesis Research Group, which offers paid consultant services to life science companies.

Author details

¹Genesis Research Group, 111 River St, Ste 1120, Hoboken, NJ 07030, USA.

Received: 25 September 2023 Accepted: 20 August 2024

Published online: 13 September 2024

References

- 21st Century Cures Act. H.R. 34, 114th Congress. 2016. <https://www.gpo.gov/fdsys/pkg/BILLS-114hr34enr/pdf/BILLS-114hr34enr.pdf>.
- FDA. Considerations for the Use of Real-World Data and RealWorld Evidence to Support Regulatory Decision-Making for Drug and Biological Products 2021. Available from: <https://www.fda.gov/media/154714/download>.
- FDA. Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry 2021. Available from: <https://www.fda.gov/media/154449/download>.
- FDA. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision Making for Drug and Biological Products 2021. Available from: <https://www.fda.gov/media/152503/download>.
- FDA. Data Standards for Drug and Biological Product Submissions Containing Real-World Data 2021. Available from: <https://www.fda.gov/media/153341/download>.
- EMA. Real-world evidence framework to support EU regulatory decision-making: Report on the experience gained with regulator-led studies from September 2021 to February 2023. Available from: https://www.ema.europa.eu/en/documents/report/real-world-evidence-framework-support-eu-regulatory-decision-making-report-experience-gained_en.pdf.
- FDA. Framework for FDA's Real-World Evidence Program 2018. Available from: <https://www.fda.gov/media/120060/download?attachment>.
- Li M, Chen S, Lai Y, Liang Z, Wang J, Shi J, et al. Integrating real-world evidence in the regulatory decision-making process: a systematic analysis of experiences in the US, EU, and China using a logic model. *Front Med*. 2021;8:8.
- Sola-Morales O, Curtis LH, Heidt J, Walsh L, Casso D, Oliveria S, et al. Effectively leveraging RWD for external controls: a systematic literature review of regulatory and HTA decisions. *Clin Pharmacol Ther*. 2023;114(2):325–55.
- Wang X, Dormont F, Lorenzato C, Latouche A, Hernandez R, Rouzier R. Current perspectives for external control arms in oncology clinical trials: analysis of EMA approvals 2016–2021. *J Cancer Policy*. 2023;35: 100403.
- Seifu Y, Gamalo-Siebers M, Barthel FMS, Lin J, Qiu J, Cooner F, et al. Real-world evidence utilization in clinical development reflected by US product labeling: statistical review. *Therapeutic Innov Regul Sci*. 2020;54(6):1436–43.
- Purpura CA, Garry EM, Honig N, Case A, Rassen JA. The role of real-world evidence in FDA-approved new drug and biologics license applications. *Clin Pharmacol Ther*. 2022;111(1):135–44.
- Jahanshahi M, Gregg K, Davis G, Ndu A, Miller V, Vockley J, et al. The use of external controls in FDA regulatory decision making. *Therapeutic Innov Regul Sci*. 2021;55(5):1019–35.
- Goring S, Taylor A, Müller K, Li TJJ, Korol EE, Levy AR, 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(2):e024895.
- Carrigan G, Bradbury BD, Brookhart MA, Capra WB, Chia V, Rothman KJ, et al. External comparator groups derived from real-world data used in support of regulatory decision making: use cases and challenges. *Curr Epidemiol Rep*. 2022;9(4):326–37.
- Gatto NM, Campbell UB, Rubinstein E, Jaks A, Mattox P, Mo J, et al. The structured process to identify fit-for-purpose data: a data feasibility assessment framework. *Clin Pharmacol Ther*. 2022;111(1):122–34.
- Concato J, Corrigán-Curay J. Real-world evidence – where are we now? *N Engl J Med*. 2022;386(18):1680–2.
- FDA. FDA approves new dosing regimen for cetuximab 2021. Available from: <https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-new-dosing-regimen-cetuximab>.
- Zhang Q, Gossai A, Monroe S, Nussbaum NC, Parrinello CM. Validation analysis of a composite real-world mortality endpoint for patients with cancer in the United States. *Health Serv Res*. 2021;56(6):1281–7.
- Ibrahim JG, Chu H, Chen M-H. Missing data in clinical studies: issues and methods. *J Clin Oncol*. 2012;30(26):3297–303.
- Jacobs EJ, Newton CC, Wang Y, Campbell PT, Flanders WD, Gapstur SM. Ghost-time bias from imperfect mortality ascertainment in aging cohorts. *Ann Epidemiol*. 2018;28(10):691–e63.
- Siannis F. Sensitivity analysis for multiple right censoring processes: investigating mortality in psoriatic arthritis. *Stat Med*. 2011;30(4):356–67.
- Carrigan G, Whipple S, Taylor MD, Torres AZ, Gossai A, Arniere B, et al. An evaluation of the impact of missing deaths on overall survival analyses of advanced non-small cell lung cancer patients conducted in an electronic health records database. *Pharmacoepidemiol Drug Saf*. 2019;28(5):572–81.
- Lesko CR, Edwards JK, Cole SR, Moore RD, Lau B. When to censor? *Am J Epidemiol*. 2017;187(3):623–32.
- Lesko CR, Edwards JK, Moore RD, Lau B. Censoring for loss to follow-up in Time-to-event analyses of composite outcomes or in the presence of competing risks. *Epidemiology*. 2019;30(6):817–24.
- Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074–102.
- Ladanie A, Schmitt AM, Speich B, Naudet F, Agarwal A, Pereira TV, et al. Clinical trial evidence supporting US Food and Drug Administration approval of novel cancer therapies between 2000 and 2016. *JAMA Netw Open*. 2020;3(11):e2024406-e.

28. Friends of Cancer Research. Considerations for Use of Real-World Evidence in Oncology 2020. Available from: https://friendsofcancerresearch.org/wp-content/uploads/Use_of_Real-World_Evidence_in_Oncology_0.pdf.
29. Freedman LS. Tables of the number of patients required in clinical trials using the log-rank test. *Stat Med*. 1982;1:121–9.
30. Rosner B. *Fundamentals of Biostatistics*. 6th ed. Belmont: Thomson Brooks/Cole; 2006.
31. Joshi M, Pustejovsky J. Simhelpers: Helper Functions for Simulation Studies. R package version 0.1.2. 2022. <https://CRAN.R-project.org/package=simhelpers>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.