



**University of Dundee**

## **Competing risk multistate censored data modeling by propensity score matching method**

Bhattacharjee, Atanu; Vishwakarama, Gajendra K.; Tripathy, Abhipsa; Rajbongshi, Bhriku Kumar

*Published in:*  
Scientific Reports

*DOI:*  
[10.1038/s41598-024-54149-y](https://doi.org/10.1038/s41598-024-54149-y)

*Publication date:*  
2024

*Licence:*  
CC BY

*Document Version*  
Publisher's PDF, also known as Version of record

[Link to publication in Discovery Research Portal](#)

### *Citation for published version (APA):*

Bhattacharjee, A., Vishwakarama, G. K., Tripathy, A., & Rajbongshi, B. K. (2024). Competing risk multistate censored data modeling by propensity score matching method. *Scientific Reports*, 14, Article 4368. <https://doi.org/10.1038/s41598-024-54149-y>

### **General rights**

Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

### **Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.



OPEN

## Competing risk multistate censored data modeling by propensity score matching method

Atanu Bhattacharjee<sup>1</sup>, Gajendra K. Vishwakarma<sup>2✉</sup>, Abhipsa Tripathy<sup>2</sup> & Bhrigu Kumar Rajbongshi<sup>2</sup>

The potential contribution of the paper is the use of the propensity score matching method for updating censored observations within the context of multi-state model featuring two competing risks. The competing risks are modelled using cause-specific Cox proportional hazard model. The simulation findings demonstrate that updating censored observations tends to lead to reduced bias and mean squared error for all estimated parameters in the risk of cause-specific Cox model. The results for a chemoradiotherapy real dataset are consistent with the simulation results.

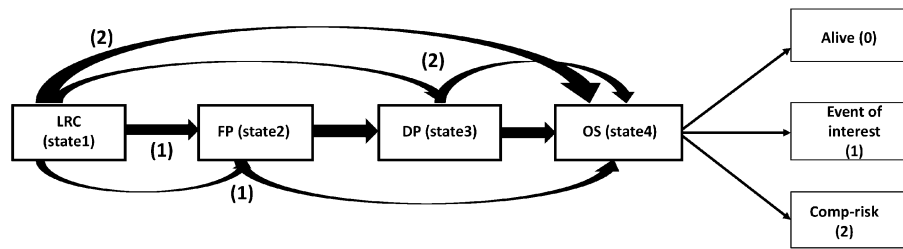
**Keywords** Multistate model, Competing risk, Non parametric estimation, Censoring, Propensity Score

Survival analysis refers to the statistical analysis of data that is measured from a particular time point until the occurrence of a specific event of interest or the attainment of a predetermined endpoint. It provides intuitive results concerning the period for events of interest, which is not confined to death but includes some other events. These events can be adverse, like the recurrence of the disease, or favorable, like recovery or discharge from the hospital. Therefore specialized techniques are devised to analyze such data with more efficacy. In cancer follow-up studies, multiple endpoints are observed, such as relapse of the disease, progression of the disease, and death status of the patient. Researchers frequently use the composite endpoint, which is defined as the time until either death or any of the nonfatal events whichever occurs first, and the information is incorporated in the final state<sup>1</sup>. In practical application, events typically arise in a sequential order, wherein the first occurrence of a disease is followed by its progression, culminating in eventual mortality. Survival time can refer to both the time survived from complete remission to recurrence and progression or overall time survived from diagnosis to death<sup>2</sup>.

In cancer studies, patients commonly go through multiple stages of the disease. In order to effectively analyse and understand the disease progression, researchers often employ the use of a multistate model (MSM) to analyse and interpret the associated data. An MSM is a continuous-time stochastic process that allows individuals to move among a finite number of states. A state can be transient if transitions emerge, and it is absorbing if no transition occurs. The primary transition pattern is modeled as a two-stage transition from “alive” to “death” state. Similarly, the illness-death model is another MSM that has been widely used in oncology studies to characterize disease development and also to investigate the rate of mortality<sup>3</sup>. However, the alive status can be further classified into two or more intermediate phases, each corresponding to different stages of illness<sup>3</sup>. The competing risk model is a particular case of MSM, which extends the basic mortality model of survival data by allowing each individual to die for numerous other reasons. Our method incorporates a multistate modeling approach, which distinguishes between individuals who have survived the nonfatal event and those who have died<sup>4</sup>. Our data contains four states, namely Loco Regional Control (LRC; state 1), First Progression (FP; state 2), Distant Progression (DP; state 3), and the final absorbing state (Death; state 4). LRC is defined as stopping the growth of cancer at its origin, hence being the state 1. Similarly, first and distant progression describes cancer progress in the local region and, thereby, in a remote region. LRC rates are often studied after providing treatments like chemotherapy to the patients. Similarly, the time duration until the final state is known as overall survival (OS). Conventionally, we consider the FP is followed by DP. Transitions are allowed from state 1 to state 2, 3, 4 progressively, however backward transition is restricted given in Fig. 1. In addition to the event-free survival probability, which has been the focus of composite endpoint analysis, the chance of being in each of these states can also be estimated<sup>5,6</sup>.

Often such data arise in oncology research when individuals are at risk of failure for  $K$  different reasons<sup>7</sup>. The study of such data necessitates a specific methodology that seeks to accurately assess the marginal probability of an event when there are several competing hazards. Conventional methods like the Kaplan–Meier survival

<sup>1</sup>Division of Population Health and Genomics, Medical School, University of Dundee, Dundee, UK. <sup>2</sup>Department of Mathematics and Computing, Indian Institute of Technology, Dhanbad, India. ✉email: vishwagk@rediffmail.com



**Figure 1.** Multi-state survival model for right censored data in presence of competing risk.

function are the most commonly used survival estimate in time-to-event data in the absence of competing events<sup>7,8</sup>. However, the use of the Kaplan–Meier estimator in the presence of competing risk most often results in an upward bias of estimation of incidence function<sup>9–11</sup>. Therefore, we require specialized techniques to account for the competing events separately. The Cumulative Incidence Function (CIF), which predicts the marginal probability for each competing event, is the most popular alternative way for analyzing competing event data. CIF can be defined as the probability of failure from a certain cause before time  $t$ , in the presence of all other causes, and can be represented as a function of individual cause-specific hazard rates. To incorporate competing events in the Cox Ph model, there are two different types of hazard functions defined for the analysis of competing risk data, such as cause-specific hazard function and sub-distribution hazard function<sup>12</sup>. The cause-specific hazard model is a statistical method used to assess the impact of covariates on the rate of occurrence of specific events of subjects who have not experienced the event of interest at the current time. On the other hand, the sub-distribution hazard ratios derived from the Fine-Gray model describe of the relative impact of covariates on the sub-distribution hazard function<sup>13</sup>. Also, the stratified hazard function can be directly modeled through regression equations based on earlier work of Gray<sup>14</sup> and Pepe<sup>15</sup>.

One challenging aspect of survival analysis is that, by the end of the study, only a few individuals experience the event, a phenomenon known as censoring. The exact survival time remains unknown for a specific group of individuals. Censoring is common in observational studies, necessitating the use of standard statistical methods to evaluate censored data, which relies on various assumptions about censoring<sup>16,17</sup>. However, if other factors influence censoring, accurate estimation of treatment effects becomes a concern, particularly when assuming that censoring is independent<sup>18</sup>. Any false assumption about censoring can lead to biased parameter estimation in the model. The survival function may be overestimated in the presence of a positive correlation between failure and censoring, or underestimated with a negative correlation<sup>18</sup>. In this study, we specifically address right-censoring cases and employ the Propensity Score Matching (PSM) technique to mitigate this issue.

Rosenbaum and Rubin proposed a methodology for assessing observational studies, wherein the likelihood of a subject belonging to a treatment (exposure) group is computed based on the covariates evaluated for that particular subject<sup>19</sup>. This score was termed as “propensity score”. The propensity score is commonly estimated through logistic regression, where the relationship between treatment (exposure) status and observable features (covariates) is examined by regression analysis. The propensity score (PS) can include a greater quantity of background covariates as it utilizes these covariates to estimate a numerical value<sup>20</sup>. After estimating the propensity score one of the methods of using this score to control the covariates is matching. In PSM a subject from the treatment group, sometimes referred to as the exposure group, is chosen randomly and subsequently paired with an untreated subject based on their propensity score<sup>19</sup>. One-to-one propensity score matching is the most prevalent method when treated and untreated subjects have similar scores<sup>21</sup>. Matching can be conducted either with or without replacement. However, matching with replacement has the potential to reduce bias and is particularly advantageous in situations where the availability of control subjects is limited<sup>22</sup>. PSM is a popular analytic method for removing the effects of confounding due to measured baseline covariates when using observational data to estimate the effects of treatment<sup>23</sup>. The propensity score method has many desirable properties like balancing the confounders between treatment and control groups to obtain conditional independence between them<sup>24</sup>. This method is used in different ways like score matching in case–control studies and Inverse Probability of Treatment Weighting (IPTW) to measure of effect frequently reported in randomized control trials<sup>25</sup>. Similarly, for observational research, Rosenbaum and Rubin proposed a propensity score method to quantify the causal impact of the treatment<sup>19</sup>. In the case of right-censored response data PSM method has been used for robust estimation of treatment effects<sup>26</sup>. In this article, we propose the use of PSM in updating individual patients’ information about censoring for different types of competing events and thereby analyzing and comparing the competing risks in the data set. The proposed method is elaborated in the manuscript.

## Methods

### Multistate model

The MSM serves as an effective method to depict complex clinical processes over time. It is a versatile framework commonly employed in clinical conditions characterized by increasing disease severity and imminent mortality. MSMs are typically assumed to follow the Markov process, aiding in the explanation of disease progression and all potential transitions. This approach also identifies previously visited states, enhancing visual representation<sup>27</sup>. While MSMs can accommodate various endpoints, our study specifically considers death as an absorbing state.

It is a continuous-time stochastic process, represented by  $X_t, t \in T$ , where an individual is allowed to transit from an initial state at  $T = 0$  to several intermediate states and finally to the absorbing state at  $T = t$ .

Let  $X(t)$ , be a finite state with state space  $S = \{1, 2, \dots, q\}$ . The process has a initial distribution defined as,  $\pi_i(0) = \text{Prob}(X(0) = i, i \in S)$ . The transition probability from state  $i$  to  $j$  is considered to be following the Markovian process and is defined as;

$$P_{ij}(s, t) = P(X(t) = j | X(s) = i) ; i, j \in S, s, t \in T \quad (1)$$

Covariates can be incorporated into the model through transition intensities in a Cox proportional hazard (Cox PH) model to explain the effect of covariates in different transitions. The Cox Ph hazard equation for the transition from  $i$  to  $j$  can be written as,

$$h_{ij}(t|\mathbf{z}) = h_{ij0}(t)\exp(\beta_{ij}\mathbf{z}') \quad (2)$$

where  $T = t$ , represents the time when the individual reaches state  $j$ , from state  $i$ .  $\mathbf{z}$  is the vector of covariates, and  $h_{ij0}$  is the baseline hazard of transition  $i \rightarrow j$ . This is also known as the transition-specific Cox PH model. Intermediate events frequently alter the path of disease progression, which results in a shift in the role of some prognostic factors following that event.

Survival data mostly contains censored information, and in this study, we'll be focusing on right censored data. For  $n$  no. of individuals let  $t_l$  be the event time and  $c_l$  be the censoring time. The primary assumption for right censored data is censoring distribution and event time distribution are independent of each other. Let  $\delta_l$  be the event indicator. Then  $\delta_l = 1$  if  $t_l \leq c_l$ , and 0 otherwise.

### Multistate with competing risk

The presence of competing events in survival data prevents the occurrence of the primary event of interest<sup>12</sup>. A MSM in the presence of competing risk has an initial state,  $q$  no. of intermediate events, and one final state which is further split into three possible events such as censored, death due to the primary event, and the competing event. Individuals may either progress to the endpoint through several intermediate stages or directly move to the absorbing state. In such cases, if an individual dies with no intermediate events in his prognosis, then such a death is termed a competing risk or competing event; otherwise, it is known as a primary event. These two events are mutually exclusive, as an individual who directly progresses to the death stage cannot experience intermediate states and vice versa.

Let the type of events at the final state be denoted by  $k = 0, 1, 2$ , where death due to local progression is denoted by  $k = 1$  and competing event denoted by  $k = 2$ . Cause-specific hazard rate describes the instantaneous rate of occurrence of the  $k$ th event in subjects who are currently event-free and have not yet experienced any type of event.

It is defined as,

$$\lambda_k(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t, D = k | T \geq t)}{\Delta t} \quad (3)$$

where  $D$  represents the type of event;  $k = 1, 2$ .

Similarly, the cumulative cause-specific hazard for individual causes is given by,

$$A_k(t) = \int_0^t \lambda_k(u) du \quad (4)$$

for  $k = 1, 2$ . The probability of occurrence of each type of event  $k$  up to a given time  $t$  is called as 'Cumulative Incidence Function' (CIF). CIF is also defined as  $1 - S(t)$ , which calculates the probability of an event occurring while considering competing risk events. The CIF of  $k$ th cause is given by,

$$F_k(t) = P(T \leq t, D = k) = \int_0^t \lambda_k(u) P(T \geq u) du = \int_0^t \lambda_k(u) [1 - A_k(u)] du \quad (5)$$

for  $k = 1, 2$ . CIF is not being calculated for  $k = 0$  as it represents censored individuals. However, as there is no one-to-one correlation between hazard and cumulative incidence, the influence on hazard cannot be immediately related to the effect on CIF<sup>28</sup>. Therefore Fine-Gray sub-distribution hazard model was developed to link covariates to cumulative incidence directly. The sub-distribution hazard model developed by Fine-Gray<sup>29</sup>. It describes the instantaneous risk of failure from  $k$ th event in subjects who are currently event-free as well as those who have previously experienced a competing event<sup>14</sup>. The **cuminc** function package **cmprsk** can estimate the CIFs from different causes of failure and allow comparison between groups.

However the competing events can be incorporated in Cox PH model as well and defined as,

$$h_k(t|\mathbf{z}) = h_{k0}(t)\exp(\beta_k\mathbf{z}) \quad (6)$$

where  $h_{k0}(t)$  is the baseline of the cause-specific hazard function and the vector  $\beta_k$  represents the covariate effect on the effect of interest.

### Propensity score matching method

Information is often incomplete due to censorship, and there are competing risk events. This means some people may have died, but we don't have their exact status. This information gap can affect comparing treatment effects. PSM is a method to create a comparable control group by matching treated and untreated individuals based on similar characteristics. We use the matching technique in PSM for individuals with incomplete survival

information. For censored individuals, PSM matches them with actual events using propensity scores based on their characteristics. Matching minimizes the impact of measured variables on the comparison of treatment effects<sup>30</sup>. Our approach involves updating individual patient information about censoring for different events and applying various models for multi-state analysis to estimate hazard rates with covariates. In our model, we have individuals with certain characteristics and consider four states: LRC, FP, DP, and Death. We use the Bayesian Cox Proportional Hazards model to obtain regression parameters for covariates. Propensity scores are calculated based on the sign of the regression parameters. We define ‘closeness’ in terms of propensity scores using a distance metric, such as the Euclidean distance. This score measures the similarity between a censored individual and actual dead individual. Another way used to explain this is Rosenbaum and Rubin’s proposed employing a caliper equal to 0.25 times the propensity score. This approach has been demonstrated to effectively mitigate 98 % of the bias resulting from measured covariates<sup>31</sup>. After calculating the propensity score, we update censored individuals based on a threshold probability. If the propensity score is high, the censored person is less likely to die in real life. This updated information helps analyze the data effectively. Censored individuals are updated to competing risk states using transition probabilities. For example, an individual may transition from being censored to competing risk 1 or 2. In summary, our proposed PSM method uses propensity scores to match individuals, update censored information, and analyze survival data with competing risks and covariates.

Let there be  $n$  individuals with  $m$  covariates denoted by  $X_1, X_2, \dots, X_m$ . The number of individuals at  $i$  th state as  $S_i$ ;  $i = \{1, 2, 3, 4\}$ . In our model, we have considered the four states to be LRC, FP, DP, and Death. Regression parameters of the  $m$  covariates  $\beta' = (\beta_0, \beta_1, \dots, \beta_m)$  are obtained through the Bayesian Cox Ph model using the Markov chain Monte Carlo (MCMC) method. Depending on the sign of the  $\beta$  value propensity score is calculated as given below;  $\mathbf{X}_{min}^T = (x_{1min}, x_{2min}, \dots, x_{mmin})$  where

$$x_{lmin} = \begin{cases} \min\{x_{11}, x_{12}, \dots, x_{1n}\}; & \text{if } \beta_l > 0 \\ \max\{x_{11}, x_{12}, \dots, x_{1n}\}; & \text{if } \beta_l \leq 0 \end{cases} \quad (7)$$

where  $l = 1, 2, \dots, m$ ;  $\beta' = (\beta_0, \beta_1, \dots, \beta_m)$  are the regression parameters obtained from Bayesian Cox PH model using Markov Chain Monte Carlo (MCMC). Using the Bayesian Cox Ph model on censored individuals at state  $i = 1, 2, 3, 4$ , we can identify the covariates that are impacting the deaths. The point  $X_{min}$  is set up in such a way that all the covariates that are contributing significantly to the hazard are placed to the right of the threshold value, and all other covariates are placed to the left.

The distance function or metrics specify the distance between points in space. The similarity between two points  $\mathbf{X}_{min}$  and  $\mathbf{X}_k$  is obtained using the different distance metric which is referred as propensity score, where  $\mathbf{X}_r$  is the covariate information of the  $r$ th individual who is censored. It is defined as the measure of similarity between a censored individual which is considered to be alive and actual dead individuals. The propensity score for  $r$  th censored individual is defined as

$$\Delta_r = d(\mathbf{X}_r, \mathbf{X}_{min}) \quad (8)$$

where  $d(\cdot)$  is distance metric for the  $m$  dimensional space. The metric  $d : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$  satisfies the following condition

1.  $d(\mathbf{X}, \mathbf{Y}) = 0$  iff  $\mathbf{X} = \mathbf{Y}$
2.  $d(\mathbf{X}, \mathbf{Y}) = d(\mathbf{Y}, \mathbf{X})$
3.  $d(\mathbf{X}, \mathbf{Y}) \leq d(\mathbf{X}, \mathbf{Z}) + d(\mathbf{Z}, \mathbf{Y})$  where  $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^m$

Euclidean distance metric has been used in this study to compute the propensity score. Using this method the distance between two  $m$ -dimensional vector  $\mathbf{X}, \mathbf{Y}$  is given by  $d(\mathbf{X}, \mathbf{Y}) = \sum_{v=1}^m w_v (x_v - y_v)^2$ . If we consider  $w_v = 1$  then it is unstandardized and standardized by S.D when  $w_v = \frac{1}{s_v^2}$ . The propensity score for  $r$ th individual who is censored using Euclidean distance metric is given by

$$\Delta_r^e = d(\mathbf{X}_r, \mathbf{X}_{min}) = \sum_{v=1}^m w_v (x_{vmin} - x_{vr})^2 \quad (9)$$

where  $\mathbf{X}_{min}^T = (x_{1min}, x_{2min}, \dots, x_{mmin})$  and covariate information for  $r$ th individual who is censored is  $\mathbf{X}_r^T = (x_{1k}, x_{2k}, \dots, x_{mk})$ .

Though censored individuals are considered to be alive, the caliper method uses the propensity score to match censored alive with real death. A threshold probability has been assumed to compare the propensity scores of censored individuals with that of actual death cases. If the propensity score is high, it is less likely that a censored person will die in real life. For a given value of threshold probability  $p$ , the death status for the censored individual gets updated as follows,

$$I_r = \begin{cases} 1; & \text{if } p > 1 - F(\Delta_r < \delta) \\ 0; & \text{otherwise} \end{cases}$$

where  $\delta$  is the  $p$  th quantile of  $\Delta_r$  and  $\Delta_r$  is the corresponding Euclidean distance metric. More is the  $p$  value from the cumulative probability, and higher is the chance that the individual is updated to status 1. After updating the censored information of the individuals and bifurcating them into dead or alive, we can apply the required functions and analyze the data accordingly.

Individuals get updated from censored cases (0) to competing risk 1, as well as competing risk 2. Let the transition be denoted by  $\tau_{0k}$ , where  $k = 1, 2$  represents the competing events.

### Estimation

Assuming the distribution of censored times to be non-informative that is parameters of event times distribution and the censored time distribution are two different sets, estimated values of the parameters are obtained by maximizing the likelihood function obtained from the risk set at the events times. The effect of covariates for transition  $i - j$ , can be modeled through the Cox PH model given by,

$$h_{ij}(t|\mathbf{z}) = h_{ij0}(t)\exp(\beta_{ij}\mathbf{z}') \quad (10)$$

Alternatively, we can write the model as,

$$h_{ij}(t) = h_{ij0}(t)\exp(\beta'_{ij}\mathbf{z}_{ij}) \quad (11)$$

where  $\mathbf{z}_{ij}$  is the vector of covariates specific to the transition  $i - j$ , specifically designed for individual based on his covariates. Let  $G$  be the set of all possible transitions  $i - j$ . The estimated value of the parameters can be obtained by maximizing the partial likelihood function as given below,

$$L(\boldsymbol{\beta}) = \prod_{ij \in G} L_{ij}(\boldsymbol{\beta}_{ij}) \quad (12)$$

Here in the presence of competing risk, for each independent cause  $k$ , the regression coefficient  $\beta_k$  can be estimated by maximizing the modified partial likelihood. The likelihood function is written as,

$$L(\beta_k|\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m) = \prod_{l=1}^n (h_{k0}(t_l)e^{\beta_k \mathbf{z}_l})^{\delta_l} e^{-H_{k0}(t_l)e^{\beta_k \mathbf{z}_l}} \quad (13)$$

where  $n$  denotes the total no. of individuals. Let  $\mathbf{Z}_{ij,k}$  denote this vector for competing event  $k$ , then the estimates of  $\beta$  vector can be obtained by maximising the generalized the Cox partial likelihood model given by,

$$L(\beta_k|\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m) = \prod_{i \rightarrow j} \prod_{d_{ij,k}=1}^2 \frac{\exp(\beta' \mathbf{z}_{ij,k})}{\sum_{l \in R_i(t_{ij,k})} \exp(\beta' \mathbf{Z}_{ij,l})} \quad (14)$$

where  $t_{ij,k}$  is either failure or censoring time for competing event  $k$ , for transition  $i - j$ .  $d_{ij,k} = 1$  if an individual has an event for transition  $i - j$  for competing event  $k$ , 0 otherwise.  $R_i(t)$  is the risk set for the set of individuals in state  $i$  at time  $t$ . The estimate of the cumulative baseline hazard using the Nelson–Aalen estimator is given by,

$$H_{ij,0}(t) = \sum_{t_{ij,k} \leq t} \frac{d_{ij,k}}{\sum_{l \in R_i(t_{ij,k})} \exp(\beta' \mathbf{Z}_{ij,l})} \quad (15)$$

For calculation purposes plugging in the non-parametric estimator for baseline and cumulative baseline hazard  $h_{k0}(t) = \frac{\delta}{\sum_{l \in R_i(t_{ij,k})} e^{\beta_k \mathbf{z}'_l}}$  in Eq. (13), we get the estimated parameters by maximizing the partial likelihood function given below

$$L(\beta_k|\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m) = \prod_{l=1}^n \left( \frac{e^{\beta_k \mathbf{z}'_l}}{\sum_{l \in R_i(t_{ij,k})} e^{\beta_k \mathbf{z}'_l}} \right)^{\delta_l} \quad (16)$$

where  $R(t)$  represents the risk set which consists of individuals at risk at event time  $t$ ,  $\delta_l$  be the censoring status of  $l$ th individual. Throughout we have assumed that  $\mathbf{z}_i$  does not include any intercept, as it is not estimable in Cox partial likelihood. Also  $\mathbf{z}_i$ , survival time  $T_i$ , censoring time  $C_i$  are independent of each other. For completely observed data maximum partial likelihood estimate (MPLE) is defined as

$$\hat{\beta} = \operatorname{argmax}_{\beta} L(\beta_k|\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m) \quad (17)$$

According to Cox, the conventional features of maximum likelihood estimation for large samples can be extended to the partial likelihood<sup>32</sup>. The derivatives of the likelihood concerning  $\beta$ , both the first and second derivatives, are consistent with those presented in Cox's work from 1972, except the risk set definition and the arguments of the independent variables<sup>33</sup>. Partial likelihood estimates of the coefficients  $\beta_k$  can be obtained using numerical methods of the Newton–Raphson type. MPLE can be computed using standard statistical software nowadays.

## Simulations

### Data simulation

Simulation studies are frequently used to assess how the existing and currently developed statistical models perform in the analysis of data<sup>34</sup>. Therefore it is necessary to generate a new dataset that closely mimics a real-life event. The exponential distribution, which assumes a constant underlying hazard function, or the Weibull distribution, which assumes a monotonically growing or decreasing hazard, are utilized and implemented with the

methodologies in the survival analysis<sup>35</sup>. Even though the Cox model is the most widely used method of survival analysis, a semi-parametric model cannot be used to simulate a dataset<sup>36</sup>. Therefore, a common approach has to be made to simplify the parametric assumptions about the distribution of event times. A general method known as the cumulative hazard inversion method allows one to simulate event times using any parametric formulation for the baseline hazard in a proportional hazards data generating model<sup>35,37</sup>. To overcome the difficulties in situations where cumulative baseline hazard is not invertible, Crowther and Lambert proposed an algorithm that nested numerical integration inside numerical root finding<sup>38</sup>. The “**simsurv**” package allows one to simulate event times from standard parametric distributions (exponential, Weibull, and Gompertz), two-component mixture distributions, or a user-defined hazard, log hazard, cumulative hazard, or log cumulative hazard function. Simulated datasets are used to compare the effects of various prognostic factors and calculate the uncertainty of model predictions, such as transition probabilities in multistate models and the impact of competing risks. We used “**mstate**” packages available in CRAN R to simulate survival data with multiple states in our study. Each simulated dataset contains 200 individuals in our research. We considered two different competing risk events and generated two other survival times corresponding to both the risks, namely  $s_1$  and  $s_2$ . Survival times were generated using the “**simsurv**” package listed in CRAN R.

Cox Ph model is widely used to analyze the impact of covariates on survival time. Therefore, to simulate survival time Cox Ph is used as given below

$$h(t) = h_0(t) \exp(x_1 \beta_1 + x_2 \beta_2 + x_3 \beta_3) \quad (18)$$

where  $h_0(t)$  is the baseline hazard. We have considered that survival time  $t$  follows Weibull distribution, now the baseline hazard takes the form  $h_0(t) = \gamma \lambda t^{\lambda-1}$ . Further, we assume  $S(t)$  to follow uniform distribution,  $U \sim U(0, 1)$ , in order to obtain the simulated survival time as,

$$t = S^{-1}(u) = H^{-1}(-\log u) \quad (19)$$

Now using the cumulative hazard function  $H(t) = \lambda t^\gamma$  in the above equation we get the simulated survival time from the Weibull distribution with parameters  $\lambda, \gamma$  as given by

$$t = \left[ \frac{-\log u}{\lambda e^{x\beta}} \right]^{1/\gamma} \quad (20)$$

The parameter values for the Cox regression for survival time  $s_1$  were chosen arbitrarily as follows  $\beta_1 = -0.5, \beta_2 = 0.01, \beta_3 = 0.02, \gamma = 1.5, \lambda = 0.1$ . Similarly, for survival time  $s_2$  the parameters values are chosen to be  $\beta_1 = -0.5, \beta_2 = 0.43, \beta_3 = 0.03, \gamma = 1.5, \lambda = 0.1$ .  $X_1, X_2, X_3$  are the covariates can be both continuous and categorical. Yet, in our simulation study, we have considered  $X_i$  to be continuous at the baseline level. A baseline covariate in the context of survival data is defined as a qualitative or quantitative variable that is measured or observed before randomization and is expected to influence the primary outcome variable to be examined. When investigating the relationship between a survival outcome and covariates, statisticians frequently consider the covariate's baseline value, which fails to take into account the relationship between survival outcome and changes in the values of the covariates. In this model, we have generated a continuous set of baseline covariates that follows the normal distribution.

Censoring time was taken to be 5 years, say. Then comparing the two survival times with censoring time, we generated the status of the individuals as 0, 1, and 2. A total of 100 such datasets are simulated. Using the Cox PH model, we calculated the  $\beta$  coefficients for each covariate and therefore calculated bias, bias, and mean score error (MSE), respectively.

### Analysis of simulated data

Our method includes a competing risk modeling approach, which discriminates between individuals who died due to cancer and those who have died due to causes other than cancer. Our data contains two competing risk events, namely death due to cancer denoted by status = 1 and death caused due by reasons other than cancer denoted by status = 2. Two separate survival times are generated for both the causes denoted by  $s_1$  and  $s_2$  respectively. Censoring time is assumed to be five units, i.e., the study ends after 5 years/months. Survival time  $s_1$  is compared with censoring time and accordingly status for  $s_1$  is generated as 0 and 1 denoted as  $st_1$ . Similarly, survival time  $s_2$  is compared with censoring time, and accordingly status for  $s_2$  is generated as 0 and 2 denoted as  $st_2$ . Then comparing  $st_1$  and  $st_2$  final status is generated as 0, 1 and 2. Applying Cox PH function separately for  $st_1$  and  $st_2$ , we can find out the hazards for each cause separately.

In an MSM with more than one intermediate event, censoring can take place in any of the states; as a result, to update such cases, we take the help of the PSM method. We assumed an arbitrary value of 0.9 for the threshold probability. If the cumulative distribution function of the propensity score value is greater than or equal to 0.9, the status of a censored individual is updated as dead. PSM method is applied to update those individuals, and further split them into dead and alive. A cause-specific Cox Ph model was used to select the effect of different prognostic factors on the propensity score. The R package “**survival**” was used for the analyses of the cause-specific Cox PH model. Each simulated data set was analyzed before and after applying the propensity score matching separately for both competing events. Results of the simulated data are shown in Table 1. Regarding competing event 1, the mean estimated regression coefficients for covariate  $x_2$  obtained before and after updation of the dataset using propensity score matching are 0.0144 and 0.0486, respectively. The MSE of the regression coefficient  $\beta_1, \beta_3$  reduced while using propensity score matching. Similarly, the bias for estimator  $\beta_3$  reduces to 0.0017 from 0.0023 after applying the PSM technique. Likewise, for competing event 2, the bias for regression coefficient  $\beta_2$  reduced from 0.0255 to 0.0054, MSE of  $\beta_1$  reduced to 0.0031 after the application of PSM. Using PSM we have proposed

Parameters	Mean	Bias	MSE	Mean	Bias	MSE
	Without propensity score method ( $\tau_{01}$ )			Without propensity score method ( $\tau_{02}$ )		
$\beta_1(x_1)$	-0.522	-0.022	0.209	-0.519	-0.019	0.059
$\beta_2(x_2)$	0.014	0.004	0.002	0.015	0.025	0.003
$\beta_3(x_3)$	0.017	0.002	0.002	0.014	-0.185	0.002
	With propensity score method			With propensity score method		
$\beta_1(x_1)$	-0.504	-0.004	0.156	-0.510	-0.010	0.003
$\beta_2(x_2)$	0.048	0.005	0.039	0.048	0.0054	0.018
$\beta_3(x_3)$	0.031	0.001	0.002	0.025	0.005	0.001

**Table 1.** Simulation results for 100 dataset consist of 200 subjects each with true parameter value  $\beta_1 = -0.5, \beta_2 = 0.01, \beta_3 = 0.02$  and  $\beta_1 = -0.5, \beta_2 = 0.43, \beta_3 = 0.03$ .

an approach that reduces the bias and MSE for estimating measures of effect in the Cox PH model in competing risk settings. Thereby we can say PSM method gives better estimates of regression coefficients.

### Analysis of chemoradiotherapy dataset

The proposed method is applied to the Chemoradiotherapy dataset<sup>39</sup>. This data frame contains information of 536 individuals followed up from 2012 to 2018. The listed variables are id, four different states, and a few covariates. The four different states are namely Loco Regional Control (LRC) as state 1, First Progression (FP) as state 2, Distant Progression (DP) as state 3, and Overall Survival (OS) as state 4 as shown in Fig. 1. The date of each event is given in the dataset, comparing the date of events with the last follow-up date status for each state is generated, and a separate column for status is generated.

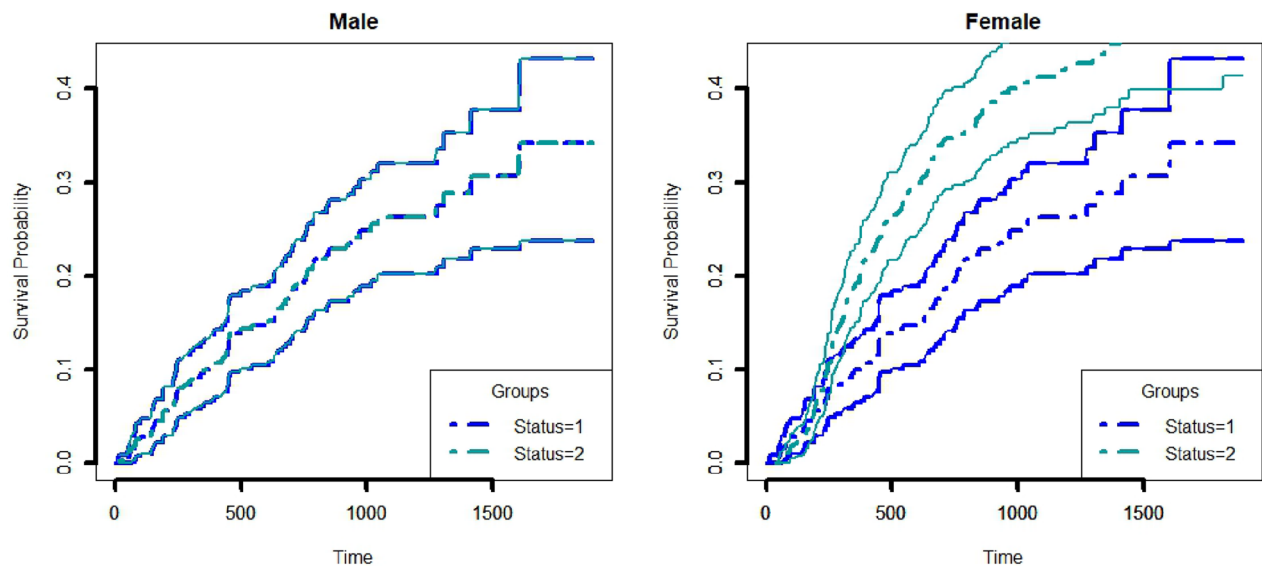
It consists of the information on causes of death due to cancer or reasons other than cancer. The covariates included in the study are baseline hemoglobin level, baseline sugar level, baseline serum creatinine, baseline serum albumin, baseline serum sodium, and baseline serum potassium denoted as cov1, cov2, cov3, cov4, cov5, and cov6, respectively. The data also provides information about the gender of each patient, which is further used in computing hazard rates for each gender separately. Individuals who progress directly from LRC to death or LRC to death without FP taking place are termed as having competing risk 2, whereas those individuals who proceed to death after a local progression taking place are classified as competing risk 1 (Fig. 1).

Cause-specific Cox proportional model is used to analyze the data. The death status of the censored cases at the final state is updated using a propensity score matching with a threshold probability of 0.9. Results of the cause-specific Cox proportional hazard model of the dataset before and after applying PSM are shown in Table 2. In the dataset, there were 154 individuals of competing risk 1, which gets updated to 156 after applying the PSM method. Similarly, for competing risk 2, 75 individuals were updated to 103 individuals. Both males and females with competing risk 1 have a higher cumulative incidence rate compared to those with competing risk 2 in the original dataset (Fig. 2). With the application of PSM, after updating the censored information difference between the CIF curves can be seen from Fig. 3, especially for female individuals, the cumulative incidence for competing risk 2 surpasses competing risk 1, for a certain period.

Parameters	$\beta$	$exp(\beta)$	$se(\beta)$	$z$	$p$	$\beta$	$exp(\beta)$	$se(\beta)$	$z$	$p$
$\tau_{01}$	Without propensity score method					With propensity score method				
cov1	-0.061	0.940	0.055	-1.100	0.271	-0.067	0.934	0.055	-1.220	0.222
cov2	0.001	1.001	0.003	0.463	0.643	0.001	1.001	0.003	0.434	0.664
cov3	-0.006	0.993	0.004	-1.623	0.104	-0.006	0.993	0.004	-1.570	0.116
cov4	-0.545	0.579	0.196	-2.778	0.005	-0.525	0.591	0.196	-2.671	0.007
cov5	-0.005	0.994	0.033	-0.155	0.877	-0.007	0.992	0.033	-0.230	0.818
cov6	-0.088	0.915	0.201	-0.401	0.659	-0.093	0.910	0.200	-0.468	0.639
Male	0.386	1.472	0.270	1.420	0.152	0.351	1.420	0.264	1.327	0.184
$\tau_{02}$	Without propensity score method					With propensity score method				
cov1	-0.115	0.890	0.076	-1.507	0.131	-0.081	0.921	0.065	-1.248	0.219
cov2	-0.000	0.999	0.005	-0.157	0.875	-0.002	0.997	0.004	-0.476	0.633
cov3	-0.011	0.988	0.006	-1.809	0.070	-0.006	0.993	0.005	-1.139	0.254
cov4	-0.731	0.481	0.252	-2.89	0.003	-0.597	0.550	0.235	-2.541	0.011
cov5	-0.101	0.903	0.042	-2.40	0.016	-0.104	0.901	0.034	-3.041	0.002
cov6	-0.291	0.747	0.285	-1.01	0.308	-0.283	0.752	0.246	-1.152	0.249
Male	0.441	1.554	0.383	1.152	0.249	0.096	1.100	0.298	0.322	0.747

**Table 2.** Results of cause specific Cox regression applied on Chemoradiotherapy data.



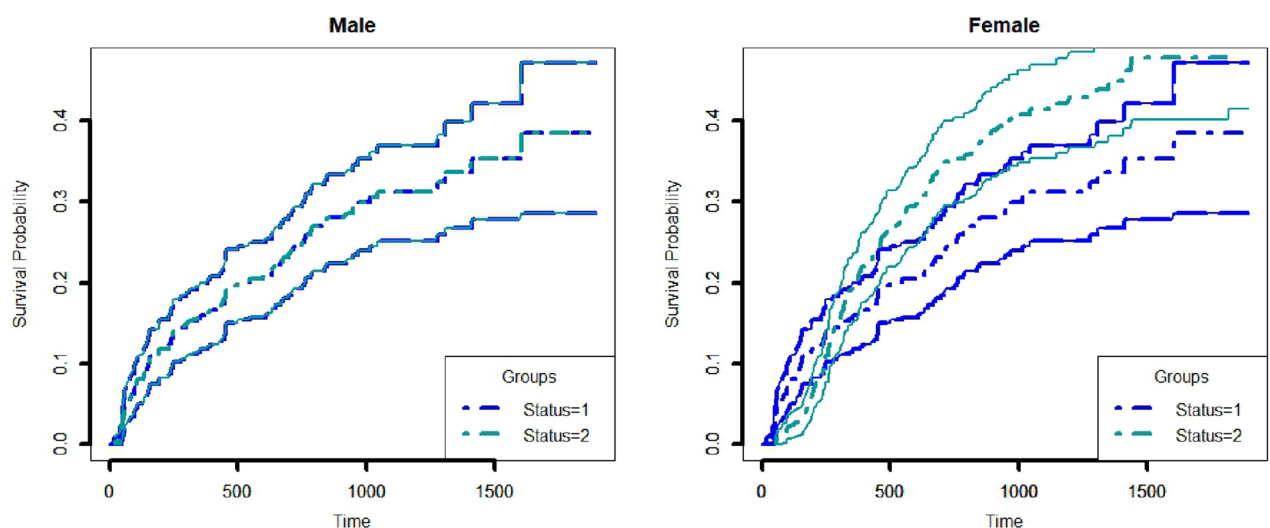


**Figure 2.** Cumulative incidence curve for male and female using nimodata. 1, 2 represents competing risks.

### Discussion

In cancer research, a variety of time-to-event endpoints can be studied, commonly known as composite endpoints<sup>40</sup>. Analysis of a single component among the competing events of time-to-event data yields more precise results<sup>41–43</sup>. The overall survival can also be considered as composite as it can be distinguished between cancer-related mortality and non-cancer mortality<sup>44,45</sup>. The intermediate events like, e.g., death after recurrence, progression, or metastasis can also be studied in a multistate framework<sup>46</sup>. For instance, the incidence of competing risk can be counted if a transition is from the initial state to the final competing event, without any progression<sup>47</sup>. Solely analysis of “progression” usually fails to account for competing risk events, which compromises the outcome<sup>48</sup>.

Survival analysis in the presence of competing events urges additional challenges for the investigator as the hazard function lacks one-to-one correspondence with the CIF<sup>28</sup>. Kaplan–Meier survival curve is not preferable to be used to estimate CIF for data with competing risk events, as it may overestimate the probability of an event. Hence Aalen–Johansen estimator is recommended. This article uses the cause-specific Cox PH model, a semi-parametric approach, to evaluate the effect of several covariates on survival after updating the censored information of individuals simultaneously. There are many such research articles on competing risk regression modeling that has been published in *Clinical Cancer Research* of late<sup>49,50</sup>. The competing risk modeling comprises of comparison of two groups of individuals who will develop the event over time with those who will not develop the event is mostly seen in various literature for time-dependent events<sup>51</sup>. However, the above statement



**Figure 3.** Cumulative incidence curve for male and female using updated nimodata by propensity score matching. 1, 2 represents competing risks.

violates the principle established by Andersen and Keiding, i.e., “Do not condition on future”<sup>52</sup>, in the analysis of survival data because it may give space to biasedness<sup>53</sup>. Within-subject cause-specific dependency for univariate and bivariate competing risk data is hardly identifiable non-parametrically, established by Tsiatis<sup>54</sup> and<sup>55</sup>. The independence assumption within the subject is hardly testable for which the validity of parametric analysis is ambiguous<sup>56</sup>. Recently, much literature has focused on cause-specific hazard function (CSH) and cumulative incidence function (CIF) to avoid practical and theoretical difficulties<sup>57</sup>. Prentice et al., in a paper, established that CSHs and CIFs are estimable for competing risk data<sup>58</sup>.

The etiology of the competing event is investigated using the cause-specific hazard function to analyze the direct association of treatment with the instantaneous hazard of the event<sup>47</sup>. Many kinds of literature on competing risk also mentioned risk-specific latent times and their association with a single cause-specific hazard, eliminating the other competing risk<sup>59</sup>. This assumption calls for the in-dependency concept of competing risk events, which is beyond the scope of the article.

When we focus our interest on the course of the disease with composite endpoints, MSM becomes very useful to explain the structure. However, the study plan is always exigent. The available approaches are listed in a few literature<sup>60</sup>. A practical application of clinical trials with diabetes patients is documented in the literature by Schulgen for sample size selection<sup>61</sup>. The appropriate size of the sample plays a vital role in the analysis of clinical data. However, such data like MSM mostly suffers from censoring of individuals. There is a sample number of individuals whose death status is censored in the dataset. However, in reality, those individuals might die due to both cancerous and non-cancerous reasons. To update the censored information, this article uses the PSM technique to update the dataset and bifurcate them into both competing events. By doing so, we have also updated the sample size in both wings of competing risk events, thereby making the analysis more meaningful. A simulation study has been carried out with simulated data for a better understanding of the impact of competing risks on the survival and efficacy of the PSM method. Similarly, the application of simulation techniques to empirical data can be seen for complex endpoint structures<sup>62</sup>. Extending the methodologies of the current research to accommodate the semi-competing risks setup would be interesting. Censored information updating using propensity score is a topic that will be investigated further in the future.

## Conclusion

In the presence of competing events in the dataset, what method to be used for analysis is determined by the type of research question. Both cause-specific hazard models and sub-distribution hazard models can be used for competing risk analysis, with model-specific advantages. This article proposes the method of propensity score matching to match the score of censored individuals with those of non-censored cases and update their status to alive or dead. Thereafter cause-specific Cox PH models have been used for competing for risk events in MSM. Also, the competing risk events are analyzed using CIF separately for the risk factors. The simulation study shows that the mean square error and bias of the regression coefficient reduced after applying PSM. The same methodology is applied to the Nimotuzumab Chemoradiotherapy dataset, and a similar result is obtained. There were 154 individuals with competing risk 1 in the dataset. When the PSM approach was used, the number increased to 156. In the same way, 75 individuals were updated to 103 for competing risk 2.

## Data availability

The data used here is not available publicly, however data can be obtained from the corresponding author on reasonable request.

Received: 24 March 2023; Accepted: 8 February 2024

Published online: 22 February 2024

## References

1. Sankoh, A. J., Li, H. & D’Agostino, R. B. Sr. Composite and multicomponent end points in clinical trials. *Stat. Med.* **36**, 4437–4440 (2017).
2. Clark, T. G., Bradburn, M. J., Love, S. B. & Altman, D. G. Survival analysis part I: Basic concepts and first analyses. *Br. J. Cancer* **89**, 232–238 (2003).
3. Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C. & Andersen, P. K. Multi-state models for the analysis of time-to-event data. *Stat. Methods Med. Res.* **18**, 195–222 (2009).
4. Eaton, A., Sun, Y., Neaton, J. & Luo, X. Nonparametric estimation in an illness-death model with component-wise censoring. *Biometrics* **20**, 20 (2021).
5. Allen, A. M. *et al.* Nonalcoholic fatty liver disease incidence and impact on metabolic burden and death: A 20 year-community study. *Hepatology* **67**, 1726–1736 (2018).
6. Holtan, S. G. *et al.* Dynamic graft-versus-host disease-free, relapse-free survival: Multistate modeling of the morbidity and mortality of allotransplantation. *Biol. Blood Marrow Transplant.* **25**, 1884–1889 (2019).
7. Scheike, T. H. & Zhang, M.-J. Analyzing competing risk data using the r timereg package. *J. Stat. Softw.* **38**, 25 (2011).
8. Klein, J. P. & Moeschberger, M. L. *Survival Analysis: Techniques for Censored and Truncated Data* Vol. 1230 (Springer, 2003).
9. Lau, B., Cole, S. R. & Gange, S. J. Competing risk regression models for epidemiologic data. *Am. J. Epidemiol.* **170**, 244–256 (2009).
10. Putter, H., Fiocco, M. & Geskus, R. B. Tutorial in biostatistics: Competing risks and multi-state models. *Stat. Med.* **26**, 2389–2430 (2007).
11. Varadhan, R. *et al.* Evaluating health outcomes in the presence of competing risks: A review of statistical methods and clinical applications. *Med. Care* **20**, S96–S105 (2010).
12. Austin, P. C., Lee, D. S. & Fine, J. P. Introduction to the analysis of survival data in the presence of competing risks. *Circulation* **133**, 601–609 (2016).
13. Austin, P. C. & Fine, J. P. Practical recommendations for reporting fine-g ray model analyses for competing risk data. *Stat. Med.* **36**, 4391–4400 (2017).
14. Gray, R. J. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *Anna. Stat.* **20**, 1141–1154 (1988).
15. Pepe, M. S. Inference for events with dependent risks in multiple endpoint studies. *J. Am. Stat. Assoc.* **86**, 770–778 (1991).

16. Leung, K.-M., Elashoff, R. M. & Afifi, A. A. Censoring issues in survival analysis. *Annu. Rev. Public Health* **18**, 83–104 (1997).
17. Berry, S. D., Ngo, L., Samelson, E. J. & Kiel, D. P. Competing risk of death: An important consideration in studies of older adults. *J. Am. Geriatr. Soc.* **58**, 783–787 (2010).
18. Huang, X. & Zhang, N. Regression survival analysis with an assumed copula for dependent censoring: A sensitivity analysis approach. *Biometrics* **64**, 1090–1099 (2008).
19. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
20. D'Agostino, R. B. Jr. Propensity scores in cardiovascular research. *Circulation* **115**, 2340–2343 (2007).
21. Austin, P. C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* **46**, 399–424 (2011).
22. Stuart, E. A. Matching methods for causal inference: A review and a look forward. *Stat. Sci. Rev. J. Inst. Math. Stat.* **25**, 1 (2010).
23. Austin, P. C. & Fine, J. P. Propensity-score matching with competing risks in survival analysis. *Stat. Med.* **38**, 751–777 (2019).
24. Cho, Y., Hu, C. & Ghosh, D. Covariate adjustment using propensity scores for dependent censoring problems in the accelerated failure time model. *Stat. Med.* **37**, 390–404 (2018).
25. Austin, P. C. The use of propensity score methods with survival or time-to-event outcomes: Reporting measures of effect similar to those used in randomized experiments. *Stat. Med.* **33**, 1242–1258 (2014).
26. Anstrom, K. J. & Tsiatis, A. A. Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data. *Biometrics* **57**, 1207–1218 (2001).
27. Hougaard, P. Multi-state models: A review. *Lifetime Data Anal.* **5**, 239–264 (1999).
28. Zhang, Z. Survival analysis in the presence of competing risks. *Ann. Transl. Med.* **5**, 14 (2017).
29. Fine, J. P. & Gray, R. J. A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **94**, 496–509 (1999).
30. Austin, P. C., Thomas, N. & Rubin, D. B. Covariate-adjusted survival analyses in propensity-score matched samples: Imputing potential time-to-event outcomes. *Stat. Methods Med. Res.* **29**, 728–751 (2020).
31. Rosenbaum, P. R. & Rubin, D. B. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am. Stat.* **39**, 33–38 (1985).
32. Cox, D. R. Partial likelihood. *Biometrika* **62**, 269–276 (1975).
33. Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B (Methodol.)* **34**, 187–202 (1972).
34. Burton, A., Altman, D. G., Royston, P. & Holder, R. L. The design of simulation studies in medical statistics. *Stat. Med.* **25**, 4279–4292 (2006).
35. Bender, R., Augustin, T. & Blettner, M. Generating survival times to simulate cox proportional hazards models. *Stat. Med.* **24**, 1713–1723 (2005).
36. Brilleman, S. L., Wolfe, R., Moreno-Betancur, M. & Crowther, M. J. Simulating survival data using the `simSurv` R package. *J. Stat. Softw.* **97**, 1–27 (2021).
37. Leemis, L. M. Variate generation for accelerated life and proportional hazards models. *Oper. Res.* **35**, 892–894 (1987).
38. Crowther, M. J. & Lambert, P. C. Simulating biologically plausible complex survival data. *Stat. Med.* **32**, 4118–4134 (2013).
39. Patil, V. M. *et al.* Low-dose immunotherapy in head and neck cancer: A randomized study. *J. Clin. Oncol.* **20**, JCO-22 (2022).
40. Manja, V., AlBashir, S. & Guyatt, G. Criteria for use of composite end points for competing risks—a systematic survey of the literature with recommendations. *J. Clin. Epidemiol.* **82**, 4–11 (2017).
41. Ryberg, M. *et al.* New insight into epirubicin cardiac toxicity: Competing risks analysis of 1097 breast cancer patients. *J. Natl. Cancer Inst.* **100**, 1058–1067 (2008).
42. Cuzick, J. Primary endpoints for randomised trials of cancer therapy. *Lancet* **371**, 2156–2158 (2008).
43. Mell, L. K. & Jeong, J.-H. Pitfalls of using composite primary end points in the presence of competing risks. *J. Clin. Oncol.* **28**, 4297–4299 (2010).
44. Braithwaite, D. *et al.* Long-term prognostic role of functional limitations among women with breast cancer. *J. Natl. Cancer Inst.* **102**, 1468–1477 (2010).
45. Yu, C.-L. *et al.* Cause-specific mortality in long-term survivors of retinoblastoma. *J. Natl. Cancer Inst.* **101**, 581–591 (2009).
46. Beyersmann, J., Allignol, A. & Schumacher, M. *Competing Risks and Multistate Models with R* (Springer, 2011).
47. Schmoor, C., Schumacher, M., Finke, J. & Beyersmann, J. Competing risks and multistate models. *Clin. Cancer Res.* **19**, 12–21 (2013).
48. Mathoulin-Pelissier, S., Gourgou-Bourgade, S., Bonnetain, F. & Kramar, A. Survival end point reporting in randomized cancer clinical trials: A review of major journals. *J. Clin. Oncol.* **26**, 3721–3726 (2008).
49. Dignam, J. J., Zhang, Q. & Kocherginsky, M. The use and interpretation of competing risks regression models. *Clin. Cancer Res.* **18**, 2301–2308 (2012).
50. Chappell, R. Competing risk analyses: How are they different and why should you care?. *Clin. Cancer Res.* **18**, 2127–2129 (2012).
51. van Walraven, C., Davis, D., Forster, A. J. & Wells, G. A. Time-dependent bias was common in survival analyses published in leading clinical journals. *J. Clin. Epidemiol.* **57**, 672–682 (2004).
52. Andersen, P. K. & Keiding, N. Interpretability and importance of functionals in competing risks and multistate models. *Stat. Med.* **31**, 1074–1088 (2012).
53. Beyersmann, J., Wolkewitz, M. & Schumacher, M. The impact of time-dependent bias in proportional hazards modelling. *Stat. Med.* **27**, 6439–6454 (2008).
54. Tsiatis, A. A nonidentifiability aspect of the problem of competing risks. *Proc. Natl. Acad. Sci.* **72**, 20–22 (1975).
55. Pruitt, R. C. Identifiability of bivariate survival curves from censored data. *J. Am. Stat. Assoc.* **88**, 573–579 (1993).
56. Wienke, A. *et al.* The heritability of cause-specific mortality: A correlated gamma-frailty model applied to mortality due to respiratory diseases in danish twins born 1870–1930. *Stat. Med.* **22**, 3873–3887 (2003).
57. Cheng, Y. & Fine, J. P. Cumulative incidence association models for bivariate competing risks data. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **74**, 183–202 (2012).
58. Prentice, R. L. *et al.* The analysis of failure times in the presence of competing risks. *Biometrics* **20**, 541–554 (1978).
59. Andersen, P. K., Abildstrom, S. Z. & Rosthøj, S. Competing risks as a multi-state model. *Stat. Methods Med. Res.* **11**, 203–215 (2002).
60. Latouche, A. & Porcher, R. Sample size calculations in the presence of competing risks. *Stat. Med.* **26**, 5370–5380 (2007).
61. Schulgen, G. *et al.* Sample sizes for clinical trials with time-to-event endpoints and competing risks. *Contemp. Clin. Trials* **26**, 386–396 (2005).
62. Allignol, A., Schumacher, M., Wanner, C., Drechsler, C. & Beyersmann, J. Understanding competing risks: A simulation point of view. *BMC Med. Res. Methodol.* **11**, 1–13 (2011).

## Author contributions

The concept was drafted by A.B. and G.K.V; A.T. and B.K.R contributed to the analysis design, and A.B. and G.K.V advised on the paper and assisted in paper conceptualization. All authors contributed to the comprehensive writing of the article. All authors read and approved the final manuscript.

### Additional information

**Correspondence** and requests for materials should be addressed to G.K.V.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>.

© The Author(s) 2024