# Analysis of Survival Data with Multiple Events

LUÍS MEIRA-MACHADO
University of Minho
Centre of Mathematics
4704-553 Braga
PORTUGAL

CARLA MOREIRA
University of Minho
Centre of Mathematics
4704-553 Braga
PORTUGAL

GUSTAVO SOUTINHO
EPIUnit-Institute of Public Health ISPUP
University of Porto
Rua das Taipas 135, 4050-600 Porto
PORTUGAL

MARTA AZEVEDO
Accenture Technology
Braga
PORTUGAL

*Abstract:* An important aim in biomedical studies is to study how an intermediate event and prognostic factors influence the course of disease of a patient. In most cases, the effect of the intermediate event is considered a time-dependent covariate and studied using extensions of the Cox proportional hazards model. Additionally, many of these studies often involve several endpoints, making the traditional approaches much more complicated. In such cases, multi-state models provide a useful tool to describe the survival process. This article aims to illustrate how multi-state models can be used as an alternative to traditional approaches. It also aims to offer guidelines for the correct use of these approaches through the analysis of survival data of patients with breast cancer. Several analyses were performed, and methods to evaluate the effect of covariates on transition intensities and to test some usual assumptions are discussed. Tree-based survival models, like the Cox proportional hazards models, are popular methods for constructing a prediction model in the field of medical research. We also present the results obtained by applying some tree-based models to the breast cancer data while showing their interpretation and utility. An overview of available software and software developed by the authors is provided to aid researchers in choosing an appropriate software tool for their purposes.

*Key-Words:* Cox proportional hazards model, machine learning methods, multi-state models, nonlinear regression, survival analysis, survival tree models

Received: April 15, 2022. Revised: October 23, 2022. Accepted: November 26, 2022. Published: December 31, 2022.

## 1 Introduction

In many clinical studies, patients may experience multiple events over a follow-up period. Traditionally, the intermediate events are considered time-dependent covariates (i.e., covariates that may change their value over time) and studied using extensions of the Cox proportional hazards model. The Cox model with time-dependent covariates can be used, but its use is more complicated in practice than the Cox model with fixed covariates. Such analysis can also be performed using multi-state models ([1], [2], [3], [4]). These models can be successfully used to investigate the progress of patients over a given number of states. Multi-state models can be illustrated graphically using diagrams with boxes representing states and with arrows between states representing possible transitions. In general, states represent the occurrence of an event that may be related to the survival prognosis, such as complications after surgery, relapses, or non-fatal episodes. In biomedical applications, states may also represent health conditions (e.g., healthy, illness, and death), or states of a disease (such as states of cancer or HIV infection). The complexity of the multi-state model depends greatly on the number of states and also on the possible transitions. The illness-death model (Figure 1) plays a central role in the theory and practice of these models, describing the dynamics of ‘healthy’ individuals who may pass into an intermediate ‘illness’ state before entering an absorbing state that in many situations is represented by the death of the individual. From now on, for simplicity’s sake, we will assume the illness-death model depicted in Figure 1.

An important goal in multi-state modeling is to evaluate the possible effect of a set of prognostic factors on the course of a disease. Several models have been used to relate individual characteristics to transition intensities. A common strategy, which allows

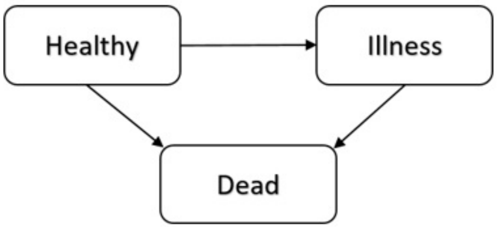

Figure 1: The progressive illness-death model.

a simplification of the analysis, consists of disaggregating the whole process into several survival models, fitting Cox regression models [5] for each of the transitions and considering some appropriate adjustments to the risk sets. In general, the model can be written as follows:

$$h_{ij}(t;X) = h_{ij,0}(t)\exp\big(\beta_{ij}^{T}X\big) \tag{1}$$

where $h_{ij,0}(t)$ denotes the baseline hazard function between states $i$ and $j$, $\beta_{ij}$ is a vector with the regression parameters, and X is a vector of covariates. For the mortality intensity function without the disease, $\alpha_{13}(t;X)$, the survival times of individuals who observed the disease are considered to be censored at the time of the disease. Individuals who remain alive and disease-free ('healthy') also contribute to censored survival times. For disease intensity, $\alpha_{12}(t;X)$, the endpoint is the time of disease onset. The survival times of individuals who did not get ill are considered censored, whether they are alive or have died unaffected by the disease. Finally, to model $\alpha_{13}(t;X)$, the mortality intensity after disease onset, only the survival times (censored or uncensored) truncated by the disease time of individuals who observed the disease are considered. Note that individuals are at risk only after entering intermediate state 2. It should be noted that in some cases, we can impose some conditions on the baseline hazard functions. For the illness-death model, an approach that is often considered is to assume that the baseline hazard functions for the transition from state 1 to state 3 ($1 \rightarrow 3$) and for the transition between state 2 and state 3 ($2 \rightarrow 3$) to be proportional. In these cases, the model for these transitions is given by:

$$\alpha_{13}(t;X) = \alpha_{13,0}(t)\exp(\beta_{13}^{T}X) \tag{2}$$

and

$$\alpha_{23}(t;X) = \alpha_{13,0}(t)\exp(\beta_{23}^{T}X + \lambda) \tag{3}$$

When implementing a Cox regression model as presented in the equations above, it is assumed that the effect of each of the continuous covariates has a linear (or log-linear) functional form. However, it should be noted that the presence of a non-linear effect can lead to serious consequences with an incorrect specification of the model, resulting in biases and a decrease in the power of statistical significance tests ([6], [7]). An incorrect functional form can also lead to a diagnosis of non-proportional hazards. The lack of flexibility of (semi-)parametric survival models has led in recent decades to the development of a variety of non-parametric regression methods based on various statistical models, of which we highlight: the Aalen additive hazards model approach [8] and Cox regression models with additive predictors [9]. To introduce flexibility into the Cox regression model, various smoothing methods can be applied, but penalized splines (P-splines), introduced by Eilers and Marx [10], are the most considered in this context. The model can be written as follows:

$$h_{ij}(t;X) = h_{ij,0}(t)\exp\Big(\sum_{k=1}^{q} f_{k,ij}(X_k)\Big) \tag{4}$$

where $f_{k,ij}(\cdot)$, $k = 1, \cdots, q$ are smooth functions associated with quantitative covariates.

In the multi-state context, two different models can be considered based on assumptions made about the dependence of transition intensities and process history. In the illness-death model this assumption is relevant on transition from the intermediate state to the absorbing state. The transition intensities can be modeled using separate Cox models, assuming that the process is Markovian (which states that the past and future are independent, given the current state of the process). Under this approach, known as 'clock forward', time $t$=0 is the start of the study and $t$ refers to the time since entry into the study. In situations where the process does not verify the Markov assumption, it is usual to use a semi-Markov model in which the future of the process is assumed to depend not on the current time but on the duration in the current state. Semi-Markov models are also called 'clock-reset' models because each time the individual enters a new state, time is reset to 0.

The Markov assumption is thus relevant in multi-state modeling and can be checked by including covariates depending on the history. In the case of the illness-death model, the Markov assumption is only relevant for the mortality transition after recurrence. We can therefore check this assumption by ascertaining whether the length of stay in the initial 'Healthy' state (i.e., the past) is important in the transition from the recurrence state to death (i.e., the future). To test this assumption in practice, let us denote by $X$ the length of stay in the initial state. Fitting a Cox regression model $\alpha_{23}(t;X) = \alpha_{23,0}(t)\exp(\beta X)$, we need to test the null hypothesis, $H_0$: $\beta = 0$, against the more general alternative, $H_1$: $\beta \neq 0$. This allows us to assess whether the intensity of the transition from the disease state to death is unaffected by the length of stay in the previous state (i.e., whether the Markov assumption is valid). Alternative methods for testing the Markov assumption in multi-state models were discussed in recent papers by Soutinho and Meira-Machado [11] and Titman and Putter [12]. The methods proposed in these two recent papers are based on measuring the discrepancy between the Aalen-Johansen estimator of the transition probabilities (a relevant predictive quantity in multi-state models) that gives consistent estimators in Markov processes, and recent approaches that do not rely on this assumption. Details on the estimation of these quantities can be seen in the paper by Meira-Machado and Sestelo [4]. To be specific, the paper by Soutinho and Meira-Machado [11] proposes a test statistic that is based on the difference of the areas under the two estimated curves of the transition probabilities. In the paper by Titman and Putter [12] a log-rank test is used on specific transitions to compare the two estimated curves. Simulation results reveal that the two recent methods perform similarly revealing high power to detect a failure of the Markov condition. Tree survival models, like the Cox proportional hazard model, can also be used for constructing prediction models. The hierarchy structure of these models makes them a good alternative because they provide simple interpretations of the relationship between covariates and hazards. Additionally, adding a new patient to these models is simple. Survival trees and survival random forests are among the most popular nonparametric alternatives to the previous methods. They make no distributional assumptions on the data, providing powerful predictive tools that offer great flexibility.

The next sections present the results of applying the methods mentioned here to a real breast cancer case carried out within the German Breast Cancer Study Group. In Section 2, we introduce the dataset and present a detailed analysis based on the Cox proportional hazards model and its extensions to cope with multiple events. An alternative analysis obtained by applying some of the most common tree-based models is given in Section 3. An overview of the available and recommended software is given in Section 4, and we conclude this paper in Section 5.

## 2 Application to a real case

Several studies have been conducted in the past decades on breast cancer. Between 1983 and 1989, four clinical trials were conducted by the 'German Breast Cancer Study Group (GBSG)', including 2746 patients with positive primary breast cancer. Details about these studies can be found in the article by Schumacher et al. [13]. In this paper, we use data from one of these trials, in which a total of 720 women with breast cancer were recruited in the period July 1984 to December 1989. The data, with complete information for 686 women, is available as part of the R software library `condSURV`. In this study, patients were followed from the date of breast cancer diagnosis until censoring or death from breast cancer. Of the total of 686 women, 171 died. Of those that died, 21 had a recorded survival time equal to the recurrence time. To deal with this, and to understand the prognostic factors associated with those individuals, they will be modeled separately when using a multi-state approach. In addition to the two event times (recurrence and death) and corresponding censoring indicator functions, a vector of covariates including age, tumor size, number of positive nodes, progesterone receptor and estrogen receptor, hormone therapy, and tumor grade are also available. A description of the variables available in the database is presented in Table 1.

Table 1: A description of the variables present in the breast cancer study.

| Variable | Description |
|---|---|
| rectime | Time to recurrence |
| censrec | Occurrence of recurrence (0: right-censored data) |
| survtime | Survival time |
| censdead | Occurrence of death (0: right-censored data) |
| age | Age at diagnosis |
| size | Tumor size (mm) |
| nodes | Number of lymph nodes involved (1-51) |
| prog_recp | Number of progesterone receptors (1 – 2380) |
| estrg_recp | Number of estrogen receptors (1-1144) |
| menopause | Statute regarding menopause (1: pre, 2: post) |
| hormone | Hormone therapy (1: yes, 2: no) |
| grade | Tumor grade (I, II e III) |

The information about a possible recurrence and the corresponding times lead one to consider this covariate as time-dependent. This covariate can be considered as a transient (intermediate) state and modeled using an illness-death model with states 'Alive and disease-free', 'Alive with recurrence' and 'Dead'. Thus, the prognostic factors for breast cancer mortality were first analyzed, the results of which are presented in Table 2. For this, simple and multiple regression models were used using the Cox proportional hazards model, considering the occurrence of recurrence as a time-dependent covariate.

By analyzing the results obtained, it can be concluded that there is a strong effect of the occurrence of recurrence on survival ($P < 0.001$), where the hazard ratio in the adjusted multiple regression model is 33.5 higher for patients with recurrence. Age, tumor size, and the number of progesterone receptors are other factors that best explain mortality according to the adjusted multiple regression model. The use of inter-

Table 2: Cox regression models with recurrence as time-dependent covariate.

| | | | Simple | | | Multivariate | | |
|---|---|---|---|---|---|---|---|---|
| Variable | | n | HR | 95%CI | p-value | HR | 95%CI | p-value |
| recurrence | | 686 | 42.299 | 25.93-<br>69.01 | <0.001 | 33.565 | 20.434-<br>55.134 | <0.001<br><0.001 |
| age | | 686 | 1.002 | 0.987-<br>1.016 | 0.836 | 1.016 | 1.002-<br>1.030 | 0.028 |
| size | | 686 | 1.021 | 1.012-<br>1.029 | <0.001 | 1.013 | 1.004-<br>1.022 | 0.007 |
| nodes | | 686 | 1.071 | 1.053-<br>1.088 | <0.001 | 1.014 | 0.991-<br>1.038 | 0.228 |
| prog_recp | | 686 | 0.993 | 0.991-<br>0.996 | <0.001 | 0.996 | 0.994-<br>0.998 | 0.001 |
| estrg_recp | | 686 | 0.998 | 0.997-<br>0.999 | 0.028 | | | |
| Menopause | | | | | | | | |
| | Pre | 290 | 1 | - | | | | |
| | Post | 396 | 1.116 | 0.821-<br>1.517 | 0.484 | | | |
| Hormone | | | | | | | | |
| | no | 440 | 1 | - | 1 | - | | |
| | yes | 246 | 0.771 | 0.559-<br>1.061 | 0.111 | 0.918 | 0.656-<br>1.285 | 0.619 |
| Grade | | | | | | | | |
| | I | 81 | 1 | - | 1 | - | | |
| | II | 444 | 3.465 | 1.522-<br>7.885 | <0.001 | 1.149 | 0.493-<br>2.678 | 0.747 |
| | III | 161 | 6.438 | 2.776-<br>14.930 | <0.001 | 1.551 | 0.639-<br>3.762 | 0.332 |

actions between the (time-) fixed covariates and the time-dependent covariate can be considered as a flexible (but less ambitious) form of multi-state modeling. In this case, where the time-dependent covariate is binary, this modeling corresponds to a situation where proportionality of hazards is assumed for transitions $1 \to 3$ and $2 \to 3$ while transition $1 \to 2$ is not modeled. The interaction between the time-dependent covariate and the fixed covariates allows modeling the situations in which the covariate has different effects before and after the time-dependent covariate occurs (intermediate event - recurrence). The results of applying this 'partial' multi-state model to breast cancer data are shown in Table 3.

Table 3: Multiple Cox regression model with interactions with recurrence.

| Variable | HR | 95%CI | p-value |
|---|---|---|---|
| recurrence | 244.26 | 8.048-7413.9 | 0.002 |
| recurrence0:age | 1.049 | 1.003-1.099 | 0.038 |
| recurrence1:age | 1.012 | 0.997-1.027 | 0.107 |
| recurrence0:size | 1.016 | 0.989-1.044 | 0.246 |
| recurrence1:size | 1.012 | 1.002-1.022 | 0.017 |
| recurrence0:nodes | 1.041 | 0.986-1.098 | 0.144 |
| recurrence1:nodes | 1.010 | 0.986-1.036 | 0.416 |
| recurrence0:prog_recp | 0.995 | 0.989-1.001 | 0.106 |
| recurrence1:prog_recp | 0.997 | 0.994-0.999 | 0.003 |
| recurrence0:hormone(s) | 0.862 | 0.348-2.131 | 0.747 |
| recurrence1:hormone(s) | 0.938 | 0.653-1.349 | 0.731 |
| recurrence0:gradeII | 0.884 | 0.191-4.097 | 0.875 |
| recurrence0:gradeIII | 1.397 | 0.268-7.299 | 0.692 |
| recurrence1:gradeII | 1.241 | 0.446-3.455 | 0.700 |
| recurrence1:gradeIII | 1.663 | 0.574-4.822 | 0.349 |

The results suggest that age has an effect on pre-recurrence survival time ($P = 0.038$) with an HR of 1.049 (95% CI: $1.003-1.099$), but also suggest that the same covariate will be of less importance in explaining the survival of individuals who recurred ($P = 0.107$). There is also statistical evidence that tumor size ($P = 0.017$) and the number of progesterone receptors ($P = 0.003$) are prognostic factors for post-recurrence survival time, with adjusted HRs of 1.012 and 0.997, respectively. It can also be seen that the remaining covariates were not identified as risk factors for the occurrence of death (with or without recurrence).

Next, we aim to study the prognostic factors not only regarding mortality (with and without recurrence) but also regarding the occurrence of recurrence. To this end, Cox models were used for each of the transitions of the multi-state model associated with breast cancer data. In the case of the models fitted for the transitions from initial state 1 to states 2 and 3 ($1 \to 2$ and $1 \to 3$), the 686 individuals who started the study were considered. For the $2 \to 3$ transition, only the 261 women who relapsed were considered. Before performing this multistate modeling (via Cox regression), it is necessary to verify the Markov assumption that the past and future of the disease depend only on the patient's current state. To this end, the influence of the time that the individual remains healthy (alive and with no evidence of disease - state 1) on the transition from the intermediate state to the absorbing state (i.e., mortality intensity in individuals who suffered from recurrence) was analyzed. From the results obtained, there is evidence that length of stay has no influence on post-recurrence survival times and, consequently, it can be assumed that a Markov model is satisfactory for the study at hand ($P = 0.121$). This conclusion was also verified ($P = 0.25$) when using the recent methods proposed by Soutinho and Meira-Machado [11]. The influence of the covariates for each of the three transitions, assuming the Markov model, is presented in Table 4 (single regression models) and Table 5 (multiple regression model). The results of the analysis of the simple Cox regression models were considered for the choice of the multiple Cox regression model.

It is noteworthy that the covariate age shows greater importance when adjusted by the multiple regression model. This was already the case when adjusting the regression model with time-dependent covariates. Conversely, at transitions $1 \to 3$ and $2 \to 3$, the covariates grade (tumor grade) and nodes (number of lymph nodes with the tumor) saw their importance decrease in the multiple regression model, partly explained by their correlation with the covariate size (tumor size). The Markov model approach and the Cox regression model approach, with recurrence as a time-dependent covariate, provide similar results, revealing the impact of age on mortality in patients without recurrence and of tumor size and progesterone receptors on mortality in patients with recurrence. The advantage of the Markov model is that it allows consideration of the effects of prognostic factors on recur-

Table 4: Markov models via simple Cox regression for all transitions.

| | | Recurrence | | | Mortality without recurrence | | | Mortality after recurrence | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | | HR | 95%CI | p-value | HR | 95%CI | p-value | HR | 95%CI | p-value |
| age | | 0.992 | 0.980-1.004 | 0.183 | 1.046 | 0.999-1.094 | 0.051 | 1.011 | 0.997-1.025 | 0.129 |
| size | | 1.014 | 1.007-1.022 | <0.001 | 1.023 | 1.001-1.047 | 0.043 | 1.010 | 1.001-1.019 | 0.039 |
| nodes | | 1.060 | 1.045-1.074 | <0.001 | 1.073 | 1.023-1.126 | 0.004 | 1.025 | 1.001-1.089 | 0.040 |
| prog_recp | | 0.997 | 0.996-0.999 | <0.001 | 0.994 | 0.988-0.999 | 0.049 | 0.996 | 0.994-0.999 | 0.001 |
| estrg_recp | | 0.999 | 0.998-1.000 | 0.056 | 0.999 | 0.995-1.002 | 0.443 | 0.999 | 0.998-1.001 | 0.232 |
| menopause | | | | 0.846 | | | 0.204 | | | 0.296 |
| | Pre | 1 | - | | 1 | | | 1 | | |
| | Post | 1.024 | 0.806-1.301 | | 1.847 | 0.716-4.765 | | 1.192 | 0.858-1.655 | |
| hormone | | | | 0.003 | | | 0.774 | | | 0.383 |
| | no | 1 | - | | 1 | | | 1 | | |
| | yes | 0.682 | 0.528-0.880 | | 0.878 | 0.363-2.127 | | 1.167 | 0.825-1.650 | |
| grade | | | | | | | | | | |
| | I | 1 | - | | 1 | | | 1 | | |
| | II | 2.527 | 1.517-4.210 | <0.001 | 1.299 | 0.291-5.809 | 0.732 | 1.760 | 0.645-4.807 | 0.270 |
| | III | 3.234 | 1.880-5.563 | <0.001 | 2.703 | 0.559-13.07 | 0.216 | 2.627 | 0.940-7.344 | 0.066 |

rence, revealing the impact of the number of nodules with tumor (nodes), status on hormone therapy, and tumor grade on recurrence.

The results by the Markov model approach indicate that except for age ($P = 0.665$) and tumor size ($P = 0.111$), the remaining covariates have an influence on the development of breast cancer recurrence. Regarding survival, for individuals in whom recurrence did not occur, only age, with an HR of 1.053, has a direct influence ($P = 0.028$). In the case of post-recurrence survival, as prognostic factors, tumor size (P = 0.030) and the number of progesterone receptors ($P = 0.004$) with HRs of 1.011 and 0.997, respectively, should be considered.

The effect of continuous covariates on the logarithm of the hazard function (log-hazards) is usually assumed to have a linear form for each of the transition intensities in a multi-state model. It turns out that this behavior is not always verified. As expected, the incorrect functional form for the covariate age led to a diagnosis of non-proportional hazards. There are numerous approaches to dealing with this problem, and the penalized spline smoothing methods (P-splines) proposed by Eilers and Marx [10] are often used in this context. These methods allow the introduction of some flexibility to the Cox model, namely in the effects of continuous covariates. The results of applying this approach to estimate the effects of continuous predictors, namely age, on the intensity of the occurrence of recurrence are presented in Table 6. Indeed, from the sample data, it is possible to observe the presence of a non-linear effect that probably would not have been detected through a parametric analysis due to the absence of prior information on the shape of the corresponding HR curve. It should be noted that the adjusted model in Table 5 (for recurrence), which considered a linear effect for continuous covariates, did not identify an effect of age on recurrence (HR=0.997; 95% CI: 0.985-1.010).

The continuous variable age was fitted with a linear and a non-linear component. The linear component has a hazard ratio function of 0.9929 ($P = 0.21$). The nonlinear component, with a probability value of 1.6e-05 indicates that the effect of age on recurrence is nonlinear. To obtain interpretable results in a simple and summary manner, we constructed flexible HR curves with 95% confidence intervals to describe the relationship between age and the (logarithm) hazard ratio for recurrence, taking a specific value as a reference. Figure 2 shows the corresponding curve for a reference value of 50 years, a value selected as a possible value for the onset of menopause. The corresponding graph confirms the existence of a non-linear effect between a woman's age and the hazard of recurrence, showing a decreasing relationship with age, with a slight increase after age 47 and remaining ap-

Table 5: Markov models via multiple Cox regression for all transitions.

| Variable | | Recurrence | | | Mortality without recurrence | | | Mortality after recurrence | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | HR | 95%CI | p-value | HR | 95%CI | p-value | HR | 95%CI | p-value |
| age | | 0.997 | 0.985-1.010 | 0.665 | 1.053 | 1.006-1.102 | 0.028 | 1.013 | 0.998-1.027 | 0.095 |
| size | | 1.007 | 0.999-1.015 | 0.111 | 1.019 | 0.992-1.046 | 0.166 | 1.011 | 1.001-1.021 | 0.030 |
| nodes | | 1.050 | 1.034-1.066 | <0.001 | 1.052 | 0.995-1.112 | 0.076 | 1.009 | 0.984-1.034 | 0.493 |
| prog_recp | | 0.998 | 0.997-0.999 | <0.001 | 0.995 | 0.989-1.001 | 0.071 | 0.997 | 0.995-0.999 | 0.004 |
| hormone | | | | | | | | | | |
| | no | 1 | - | | 1 | | | 1 | | |
| | yes | 0.717 | 0.551-0.932 | 0.013 | 0.826 | 0.334-2.038 | 0.677 | 0.957 | 0.665-1.378 | 0.814 |
| grade | | | | | | | | | | |
| | I | 1 | - | | 1 | | | 1 | | |
| | II | 2.034 | 1.214-3.407 | <0.001 | 0.907 | 0.197-4.179 | 0.900 | 1.215 | 0.436-3.387 | 0.710 |
| | III | 2.269 | 1.301-3.955 | <0.001 | 1.602 | 0.302-8.068 | 0.576 | 1.593 | 0.558-4.630 | 0.392 |

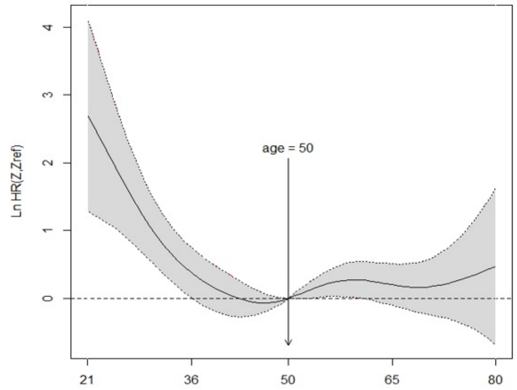

Figure 2: Nonparametric estimates of hazard rate recurrence dependence (with 95% confidence limits) in breast cancer patients. Reference value of 50 years of age.

proximately constant thereafter. The data is quite dispersed at older ages, as reflected by the wide confidence interval at these ages. This graphical representation reveals that the hazard of recurrence is higher for younger women. For example, the hazard ratio function takes the value of $\exp(1.1032) = 3.0138$ (with 95% CI 1.8083-5.0230) when a 30-year-old patient is compared with a 50-year-old patient (reference value).

Table 6: Multiple Cox regression model for the intensity of recurrence occurrence with a non-linear effect for age.

| Variable | | HR | 95%CI | p-value |
|---|---|---|---|---|
| age | | | | |
| | age (Linear) | 0.993 | | 2.1e-01 |
| | ps (age,df=4.9) (nonlinear) | | | 1.6e-01 |
| nodes | | 1.047 | 1.032-1.063 | 6.9e-10 |
| size | | 1.009 | 1.001-1.017 | 3.1e-02 |
| prog_recp | | 0.998 | 0.997-0.999 | 4.4e-04 |
| Hormone | | | | 3.7e-03 |
| | no | 1 | - | |
| | yes | 0.675 | 0.518-0.880 | |
| Grade | | | | |
| | I | 1 | - | |
| | II | 1.960 | 1.169-3.287 | 1.0e-02 |
| | III | 2.078 | 1.188-3.634 | 3.7e-03 |

Finally, it is worth mention that the proportional hazards assumption has been verified for all multivariable Cox models with the exception of the multivariable Cox model for recurrence with age as a linear effect (Table 5).

## 3 Tree-based models

The Cox proportional hazards model is the most common tool for studying the effect of predictor variables on survival. However, this model assumes some known assumptions that may not be verified in some databases. Among these assumptions, the proportional hazards assumption and the log-linear relationship between the independent covariates and the underlying hazard function stand out. For these cases, more flexible nonparametric approaches are desired. Survival trees and survival random forests are among the most popular nonparametric alternatives to the Cox proportional hazard model.

Tree-based methods were first introduced by Morgan and Sonquist [14] and later developed by Breiman et. al. ([15], [16]) using the CART algorithm, which stands for Classification And Regression Trees. The CART algorithm is the most widely used for constructing these models, which have been proven to be powerful predictive tools that offer great flexibility. They can detect and have in consideration nonlinear effects and detect certain types of interactions between covariates without the need to specify them in advance. Two important steps in the CART algorithm are splitting and pruning, for which several methods exist. In fact, most of the proposed tree-based methods for censored data are distinguished by the method that is used for the splitting criterion for which the log-rank test has been used by Hothorn et al. [17]. The log-rank test criterion is based on the maximization of separation between child nodes, being used to select the covariate that enter at each node as well the corresponding cutpoint. Other criteria have been proposed that aim to minimize the hazard within the child nodes. Comparative research studies of splitting methods can be found in the papers by Radespiel-Tröger et al. ([18], [19]). The survival tree is then built by recursively dichotomizing the sample. As the number of tree nodes increases, and dissimilar cases become separated, each node in the tree becomes homogeneous. The tree may grow until terminal nodes are populated by individuals with similar behavior or some stopping rule is obtained. Pruning may be used to reduce or select an optimal size of the tree by removing parts of the tree that do not provide power to classify the occurrences. To this end, the cost complexity measure, described by Breiman [15], can be used. A good review of the topic that includes a method for constructing the survival tree can be found in the paper by Bou-Hamad et al. [20].

The obtained tree structure for recurrence intensity, mortality without recurrence and for time from recurrence to death are given in Figures 3, 4, and 5, respectively. The circles in the figures represent the three internal nodes, whereas Kaplan-Meier plots of survival are shown at the terminal nodes. For each of the internal node it is shown the covariate used to split the sample. Probability values obtained using the log-rank test according to the splitting value can also be obtained. All p-values of the log-rank test were lower than 0.05. The number of subject in the terminal nodes are also given. The covariates used in the tree structure for the recurrence transition were, nodes and progesterone receptors. The first split is based on the variable node: subjects with a value lower or equal to 3 go to the left node, and those with a value greater than 3 go to the right node. After the first split, we can see that the subjects with a value of nodes greater than tend to have a lower survival time. Among those, subjects with a value of progesterone receptors lower or equal to 55 have the worst survival prognosis. As shown in Figures 4 and 5, progesterone receptors are the best predictors for survival (with and without recurrence), being used to split the sample in the first node of the tree, whereas the second split is based on positive nodes. Again, the Kaplan-Meier survival estimates for the terminal nodes show that they are all different from each other and provide useful information that can be used for prediction purposes. For example, it can be seen that subjects with progesterone receptors lower or equal to 35 have a lower survival time. Predictor groups can be easily obtained from these three survival trees.

It is well known that, in some cases, the use of a

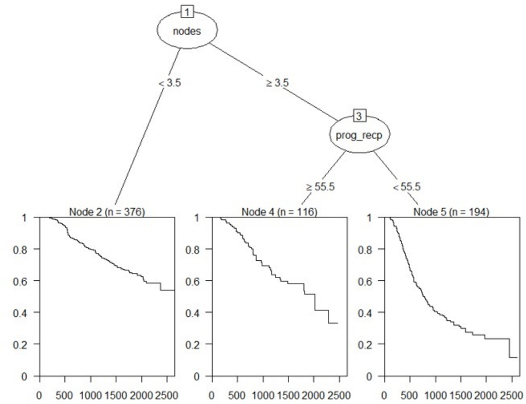

Figure 3: Survival tree for the breast cancer data. Recurrence transition.

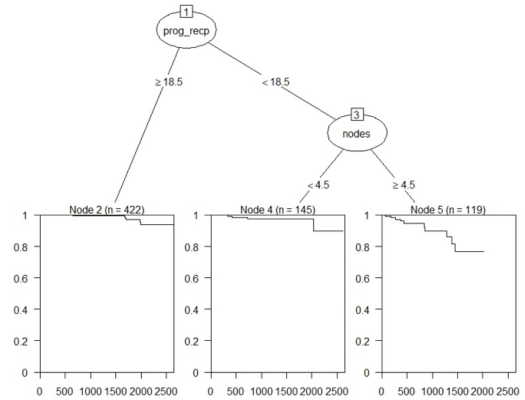

Figure 4: Survival tree for the breast cancer data. Mortality transition without recurrence.

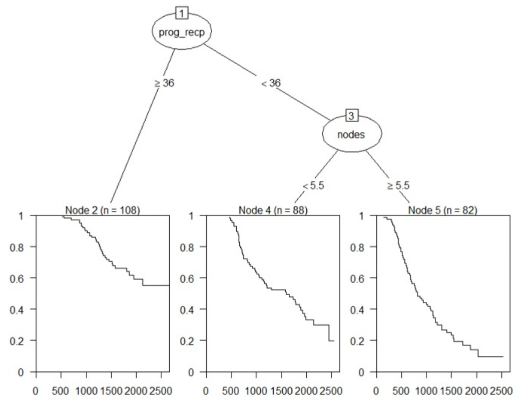

Figure 5: Survival tree for the breast cancer data. Mortality transition after recurrence.

single tree may not lead to such a good predictor of survival since its performance may reveal high variability. Aggregation methods can be used to solve this problem. Random survival forests ([21], [22], [23], [24]) are one of these approaches that basically aggregates the information from many trees, therefore providing more stable results with less variability than those obtained from a single tree. In a Random Survival Forest (RSF), randomization is first used to select a bootstrap sample that is used for growing the tree, and later it is used to select the covariate to be used to split the nodes. Predictions are formed by aggregating (averaging) predictions of individual trees in the ensemble. Besides the good ability of the RSF for predictive issues, they can also be used for studying the importance of the covariates in the prognosis. Table 7 shows the variable importance (×100) for the three breast cancer datasets. All reported values were averaged over 100 independent runs. Large importance values indicate variables with predictive ability, whereas negative or near-zero values identify nonpredictive variables to be filtered. From this table, we can conclude that positive nodes, progesterone receptors, age, and estrogen receptors are highly predictive factors of recurrence. On the other hand, menopause and tumor size are unlikely to be predictive. It can also be seen that progesterone receptors and positive nodes are good predictors for both of the two mortality transitions (with and without recurrence). Some of these findings are in agreement with the results obtained by the Cox models shown in Table 5. It is also worth mentioning the predictive effect of age on the recurrence transition, which is also in agreement with results shown in Table 6.

## 4 R packages

Several researchers have developed software for multi-state survival data analysis in recent years. A comprehensive list of packages available on the R software distribution network (CRAN) can be obtained in the 'CRAN task view' under the topic 'Survival Analysis' [25]. To provide biomedical researchers with a user-friendly tool in the context of multi-state models, a package has been developed for the R statistical software called `survidm` ([26], [27]). This package can be used to perform multi-state regression using Cox (semi-) Markov models. The package goes far beyond multi-state regression, providing biomedical researchers with the ability to obtain other interpretable results in a simple and summarized way. This includes estimates of various predictive probabilities, such as transition probabilities, occupancy probabilities, cumulative incidence functions, and the distribution of length of stay in each state. A limitation of the `survidm` package is that it can only be used for the progressive illness-death model. However, this turns out to be an advantage for users who wish to analyze data from a model with this structure. In these cases, the survidm package is ideal, as it is easy to use with a strong resemblance to the survival package, which is well known and widely used in the R software user community. For models with a structure other than the illness-death model, we recommend using the `mstate` package. These two R packages as well as the `markovMSM` package can also be used to check the markov assumption.

Table 7: Variable importance ($\times 100$) for breast cancer datasets. All reported values averaged over 100 independent runs. Each run based on 500 trees under log-rank splitting.

| Variable | Recurrence | Mortality without | Mortality after |
|---|---|---|---|
| Variable | Recurrence | recurrence | recurrence |
| nodes | **6.544878** | **2.642534** | **1.631170** |
| prog_recp | **2.309012** | **5.914801** | **6.742094** |
| age | **1.919673** | -0.889223 | -0.419527 |
| estrg_recp | **1.235000** | **1.068243** | 0.409769 |
| hormone | 0.669168 | -0.247917 | 0.200952 |
| grade | 0.569283 | -0.077202 | **1.297796** |
| size | 0.248867 | **2.851047** | 0.333358 |
| menopause | -0.003882 | -0.160168 | -0.067455 |

The `survival` and `mgcv` packages by Terry Therneau and Simon Wood, respectively, can be used to introduce flexibility into the Cox regression model whereas `CatPredi` can be used to categorize continuous predictors. The `smoothHR` library, developed by Araújo and Meira-Machado, allows the calculation of point estimates for the hazard function ratio and their corresponding confidence limits for continuous predictors considering a non-linear effect, introduced using penalized splines.

Several R packages have been recently developed that implement tree-based models. Tree-structured models for survival analysis are now implemented in `rpart` R package. Recursive partitioning algorithms have been also implemented in the `partykit` R package whereas `randomForestSRC` implements a unified treatment of Breiman's random forests [15] for survival, regression and classification problems [21].

## 5 Discussion

Multi-state models can be successfully used for describing complex event history data, for example, stages in the disease progression of a patient. Although the relevance of these models is well acknowledged, they are still not frequently applied, in particular by non-statisticians. One major goal in survival studies is to study the relationship between the different covariates and disease evolution, for which the Cox proportional hazards model is the most commonly used tool. Besides assuming proportional hazards, the effect of a given continuous predictor through the Cox model on log-hazards is modeled linearly. This paper aims to review some of the statistical methods for analyzing data with multiple events, providing alternative approaches with an emphasis on practical issues such as the estimation of the effect of covariates and how to overcome the aforementioned difficulties. We attempted to propose a comprehensive approach to analyze a variety of methodologic issues related to the analysis of survival data with multiple events using data from breast cancer patients. First, we focused on estimating, and adjusting for the impact of recurrence on survival through the use of time-dependent Cox regression analyses. The use of interactions between the fixed covariates and recurrence (time-dependent covariate) was used to obtain a flexible (but less ambitious) form of multi-state modeling. We have shown that an alternative multi-state approach that consists on using a stratified Cox regression model based on the 'clock forward' approach may be preferable. We have shown that in all these approaches, the erroneous assumption of linearity may have serious consequences. Fitting the incorrect functional effect of a continuous predictor may lead to bias and decreased power of tests for statistical significance. It may also lead to nonproportional hazards. All these issues may be corrected but they can also be avoided using popular nonparametric alternatives to the Cox proportional hazard model that are based on survival trees and survival random forests. Another interesting and possibly undervalued aspect of multi-state models is the possibility of applying these models to obtain predictions probabilities of the clinical prognosis, such as the occupation probabilities, the transition probabilities ([26], [27]), and the cumulative incidence functions. This can be seen in the paper by Meira-Machado and Sestelo [4]. All these quantities can be obtained using software developed by the authors ([28], [29]). Data and all R input commands used in this study are available upon request from the corresponding authors.

**Acknowledgements:** This work received financial support from the Portuguese Foundation for Science and Technology ("Fundação para a Ciência e a Tecnologia"), references UIDB/00013/2020, UIDP/00013/2020 and EXPL/MAT-STA/0956/2021.

**Conflict of interest** The authors declare there is no conflict of interest.

*References:*

[1]. P. K. Andersen, O. Borgan, R. D. Gill, N. Keiding, Statistical Models Based on Counting Processes, Springer-Verlag, New York, 1993. http://dx.doi.org/10.1007/978-1-4612-4348-9

[2]. P. Hougaard, Analysis of multivariate survival data, Springer, New York, 2000. https://doi.org/10.1093/ije/30.4.909

[3]. L. Meira-Machado, J. de Uña-Álvarez, C. Cadarso-Suárez, P. K. Andersen, Multistate models for the analysis of time to event data, Statistical Methods in Medical Research, 18 (2009), 195-222. https://doi.org/10.1177/0962280208092301

[4]. L. Meira-Machado, M. Sestelo, Estimation in the progressive illness-death model: a nonexhaustive review, Biometrical Journal, 61(2) (2019), 245-263. https://doi.org/10.1002/bimj.201700200

[5]. D. R. Cox, Regression models and life tables, Journal of the Royal Statistical Society Series B, 34 (1972), 187-220. https://doi.org/10.1007/978-1-4612-4380-9_37

[6]. C. A. Struthers, J. D. Kalbfleisch, Misspecified proportional hazard models, Biometrika, 73(2) (1986), 363-369. https://doi.org/10.1093/biomet/73.2.363

[7]. G. L. Anderson, T. R. Fleming, Model misspecification in proportional hazards regression, Biometrika, 82 (1995), 527-541. https://doi.org/10.2307/2337531

[8]. T. Martinussen, T. H. Scheike, Dynamic regression models for survival data, Springer, New York, 2006.

[9]. T. Hastie, R. Tibshirani, Generalized additive models, Chapman and Hall, 1990. https://doi.org/10.1201/9780203753781

[10]. P. H . C. Eilers, B. D. Marx, Flexible smoothing with B-splines and penalties, Statistical Science, 11 (1996), 89-121. https://doi.org/10.1214/ss/1038425655

[11]. G. Soutinho, L. Meira-Machado, Methods for checking the Markov condition in multi-state survival data, Computational Statistics, 37 (2022), 751-780. https://doi.org/10.1007/s00180-021- 01139-7

[12]. A. C. Titman, H. Putter, General tests of the Markov property in multi-state models, Biostatistics, 23(2) (2022), 380-396. https://doi.org/10.1093/biostatistics/kxaa030

[13]. M. Schumacher, G. Bastert, H. Bojar, K. Hiibner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R. L. A. Neumann, H. F. Rauschecker, for the German Breast Cancer Study Group (GBSG). A randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients, Journal of Clinical Oncology, 12 (1994), 2086-2093. https://doi.org/10.1200/JCO.1994.12.10.2086

[14]. L. Breiman, J. H. Friedman, R. A. Olshen, C. Stone, Classification and Regression Trees, California: Wadsworth, 1984. https://doi.org/10.1201/9781315139470

[15]. L. Breiman, Random Forests, Machine Learning, 45 (2001), 5-32. https://doi.org/10.1023/A:1010933404324

[16]. J. N. Morgan, J. A. Sonquist, Problems in the analysis of survey data, and a proposal, Journal of the American Statistical Association, 58 (1963), 415-434. https://doi.org/10.1080/01621459.1963.10500855

[17]. T. Hothorn, K. Hornik, A. Zeileis, Unbiased recursive partitioning: A conditional inference framework, Journal of Computational and Graphical Statistics, 15 (2006), 651-674. https://doi.org/10.1198/106186006X133933

[18]. M. Radespiel-Tröger, T. Rabenstein, H. T. Schneider, B. Lausen, Comparison of tree-based methods for prognostic stratification of survival data, Artificial Intelligence in Medicine, 28 (2003), 323-341. https://doi.org/10.1016/s0933- 3657(03)00060-5

[19]. M. Radespiel-Tröger, O. Gefeller, T. Rabenstein, T. Hothorn, Association between split selection instability and predictive error in survival trees, Methods of Information in Medicine, 45(5) (2006), 548-556. https://doi.org/10.1055/s- 0038-1634117

[20]. I. Bou-Hamad, D. Larocque, H. Ben-Ameur, A review of survival trees, Statistics Surveys 5 (2011) 44-71. https://doi.org/10.1214/09-SS047

[21]. H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer, Random survival forests, The Annals of Applied Statistics, 2(3) (2008) 841–860. https://doi.org/10.1214/08-AOAS169

[22]. H. Ishwaran, U. B. Kogalur, X. Chen, A. J. Minn, Random survival forests for high-dimensional data, Stat Anal Data Min, 4 (2011) 115-132. https://doi.org/10.1002/sam.10103

[23]. H. Ishwaran, U. B. Kogalur, Random survival forests for R, R News 7 (2007) 25-31.

[24]. H. Ishwaran, T. A. Gerds, U. B. Kogalur, R. D. Moore, S. J. Gange, B. M. Lau, Random survival forests for competing risks, Biostatistics 15 (2014) 757-773. https://doi.org/10.1093/biostatistics/ kxu010

[25]. A. Allignol, A. Latouche, CRAN Task View: Survival Analysis. Version 2022-03-07, URL http://CRAN.Rproject.org/view=Survival

[26]. G. Soutinho, L. Meira-Machado, Parametric landmark estimation of the transition probabilities in survival data with multiple events, WSEAS TRANSACTIONS on MATHEMATICS , 1 (2022), 207-217. https://doi.org/10.37394/23206.2022.21.27

[27]. A. Araújo, L. Meira-Machado, J. Roca- Pardiñas, TPmsm: Estimation of the Transition Probabilities in 3-State Models, Journal of Statistical Software, 62(4) (2015), 1-29. https://doi.org/10.18637/jss.v062.i04.

[28]. L. Meira-Machado, M. Sestelo, G. Soutinho, survidm: Inference and Prediction in an Illness- Death Model. R package version 1.3.2, URL https://CRAN.R-project.org/package=survidm

[29]. G. Soutinho, M. Sestelo, L. Meira-Machado, survidm: An R package for Inference and Prediction in an Illness-Death Model, The R Journal 13:2 (2021) 70-89. https://doi.org/10.32614/RJ- 2021-070

## Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)