Applied Intelligence (2022) 52:794–807
https://doi.org/10.1007/s10489-021-02311-8

Cure models to estimate time until hospitalization due to COVID-19

A case study in Galicia (NW Spain)

Maria Pedrosa-Laza[1] · Ana López-Cheda[2] · Ricardo Cao[2,3]

Accepted: 3 March 2021 / Published online: 12 May 2021

Abstract

A short introduction to survival analysis and censored data is included in this paper. A thorough literature review in the field of cure models has been done. An overview on the most important and recent approaches on parametric, semiparametric and nonparametric mixture cure models is also included. The main nonparametric and semiparametric approaches were applied to a real time dataset of COVID-19 patients from the first weeks of the epidemic in Galicia (NW Spain). The aim is to model the elapsed time from diagnosis to hospital admission. The main conclusions, as well as the limitations of both the cure models and the dataset, are presented, illustrating the usefulness of cure models in this kind of studies, where the influence of age and sex on the time to hospital admission is shown.

Keywords Censored data · COVID-19 · Hospital demand · Forecasting · Survival analysis

1 Introduction

1.1 Survival analysis

Survival analysis is the branch of Statistics which considers the study of the elapsed time until the occurrence of an event of interest [1]. Frequently, such event is death by a pathology, and thus this variable receives the name of "lifetime", and the event is called "failure" or "death".

From a statistical point of view, the survival function at time t is conceived as the probability of an individual living beyond that time. As a result, the basis of survival analysis relies on the estimation of such a probability for any value of t. Mathematically, this survival function is defined as

$$S(t) = P(T > t),$$

where T represents the "lifetime". It is important to highlight that $S(t)$ can take any shape which satisfies the following conditions [2]:

- $S(t)$ is a decreasing function
- $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$

A key concept in survival analysis is the hazard function, $h(t)$, which represents the instantaneous failure rate for a certain individual. That is, this function represents the probability that a subject will experience an event of interest within a small specific time interval, given that the individual has survived until the beginning of this interval, defined by:

$$h(t) = \lim_{\Delta t\to 0} \frac{P[t \le T < t + \Delta t | T \ge t]}{\Delta t}.$$

Note that if T is a continuous variable, $S(t)$ and $h(t)$ are directly related through

$$h(t) = -\frac{d \ln[S(t)]}{dt}.$$

This article belongs to the Topical Collection: *Artificial Intelligence Applications for COVID-19, Detection, Control, Prediction, and Diagnosis*

✉ Maria Pedrosa-Laza
maria.pedrosa@api.uniovi.es

Ana López-Cheda
ana.lopez.cheda@udc.es

Ricardo Cao
rcao@udc.es

[1] Área de Proyectos de Ingeniería, Escuela Técnica Superior de Minas, University of Oviedo, Oviedo, Spain

[2] Research Group MODES, CITIC, University of A Coruña, 15071, A Coruña, Spain

[3] ITMATI, A Coruña, Spain

A wide variety of inference techniques can be applied to estimate the survival function and, due to its direct relationship, for the hazard function. The idea is to use samples extracted from a specific population, in order to assess the behavior of the time-to-event variable. Furthermore, these analyses can be useful for the comparison of the survival curves estimated from different populations, or for the assessment of the influence of certain explanatory variables on the lifetime of a group [3].

1.2 Censored data

There are some limitations regarding data availability when performing a survival study. In some cases, the study design does not allow for the accurate measure of the lifetime for all the individuals within the sample, and this phenomenon is known as censoring. As an illustrative example, let us define the population of "patients diagnosed with a terminal disease", where the variable of study would be "elapsed time from the diagnosis until death". Due to specific circumstances that may happen during the follow-up period, such as hospital transfer, leaving the study prior to experiencing the event or premature end of study, the lifetime of some individuals will be unknown. In such cases, these observations are said to be censored.

Censoring plays a key role in survival analysis, and its influence needs to be specifically considered. We can find two major types of censoring [4]:

- Point censoring, which arises when the event of interest does not occur within the study period, known as right censoring, or when we cannot determine the exact lifetime even if the event of interest happens in between the limits of this period. In the previous example, an observation of a certain individual would be right-censored if death occurs after the end of the study, or if there is a loss of follow-up during the study. On the other hand, the observation will be left-censored if death occurs in the follow-up period but we cannot determine the date of diagnosis, previous to the beginning of the study.
- Interval censoring, which appears in these cases when the event happens between two exact time points, but it is not possible to determine the exact point of occurrence within the interval. For example, if the study variable is "time to recurrence" of a certain disease, the data will be interval censored if a patient does not suffer from the disease in a follow-up visit, but he or she does present it in the following medical check-up, being thus impossible to determine the exact point between both visits when the symptomatology has appeared.

This paper focuses on right censoring, which is the most common censoring case in clinical studies [4]. Therefore, in order to deal with censoring, some specific notation needs to be introduced [5]. The random variable "real lifetime", even though it is not always observed due to censoring, is represented by Y. Its probability density function is denoted by $f(y)$ and its survival function by $S(y)$.

The maximum lifetime that can be observed for each individual, due to the aforementioned limitations, is known as "censoring time" and denoted by C. Its value is determined by the end of the study, or by the moment of the loss of follow-up, and can be different for each individual. Therefore, the variable Y can be observed only if condition $Y \leq C$ is fulfilled. Otherwise, the observation is censored and only C is observed. The random variable "observed lifetime", T, is defined as $T = \min(C, Y)$. If the observation is not censored, the observed lifetime will be equal to the real lifetime. Moreover, δ is the uncensoring indicator:

$$\delta = \mathbf{1}(Y \leq C). \tag{1}$$

As defined by (1), δ is equal to 0 if the observation is censored, and it is equal to 1 otherwise. The sample is just a collection of independent observations $(T_1, \delta_1), \ldots, (T_n, \delta_n)$ with the same distribution as the random variables (T, δ).

Data analysis tools for the survival curve consider censoring. One of the most popular estimators is the Kaplan-Meier (KM) estimator by [6], named after the researchers who developed it. The KM estimator is a nonparametric estimator, and thus it does not make any assumption with regards to the specific form of the probability distribution of Y. This method considers that the probability of surviving beyond time t from the beginning of the study equals the product of the n survival rates in the period $[0, t]$ [7]. Mathematically, it is defined as:

$$\hat{S}(t) = \prod_{i:T_{(i)} \leq t}^{n} \left(1 - \frac{\delta_{[i]}}{n - i + 1}\right). \tag{2}$$

In (2), $\delta_{[i]}$ is the corresponding uncensoring indicator concomitant of $T_{(i)}$, and $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(n)}$ are the ordered observed lifetimes. It has been proved that this is the nonparametric maximum likelihood estimator for $S(t)$ [8]. Furthermore, a generalization of KM has been proposed in such a way that it allows considering the effect of a certain covariate X in $S(t)$. This generalization, introduced by [9], is known as the Beran estimator, defined as:

$$\hat{S}_h(t|x) = \prod_{i:T_{(i)} \leq t} \left(1 - \frac{\delta_{[i]} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)}\right), \tag{3}$$

where

$$B_{h(i)}(x) = \frac{K_h(x - X_{[i]})}{\sum_{j=1}^{n} (K_h(x - X_{[j]}))}. \tag{4}$$

In (4), K_h represents the rescaled kernel function with a smoothing parameter, h, and $X_{[i]}$ is the covariate concomitant of $T_{(i)}$. A suitable choice of h is critical in kernel estimation, and several bandwidth selection methods can be considered, such as bootstrap [10], plug-in [11], or cross-validation [12].

1.3 Survival analysis versus cure models

Classical survival analysis assumes that all the individuals will experience the event of interest. However, in some cases, a fraction of the population will never experience failure. This fact is especially relevant in cancer studies: when the event of interest is the recurrence of a tumor and the study variable is the "time from remission to relapse", some individuals will never suffer from a new tumor, even after a long time period. These individuals are said to be cured. Note that in a general survival analysis framework *cure* differs from the classical meaning of absence of illness, but refers to the fact that these patients are exempt to suffer the event of interest.

The application of cure models has not been limited to Medical Sciences or Epidemiology, but has also been extended to different areas such as Social Sciences, Economy or Engineering. For instance, the elapsed time from marriage to divorce, the lifespan of a specific product, or the unemployment period until an individual gets a job are all considered as time-to-event studies [13, 14].

From a practical point of view, distinguishing between cured individuals and censored observations which are susceptible to experience the event of interest is not trivial. Cure models handle this situation, becoming essential statistical techniques in cases where applying a classical survival analysis is not appropriate.

The aim of this work is to review the distinct types of cure models, comparing their strengths and limitations. Furthermore, the different approaches will be applied to a COVID-19 dataset, studying the elapsed times from the diagnosis until hospitalization.

2 Cure models

The main idea of cure models is the willingness to complete the unavailable information in order to identify and estimate the fraction of cured individuals [15]. In this case, the survival function for the population does not fulfill one of the aforementioned assumptions which hold in classical survival analysis, since $\lim_{t\to\infty} S(t) > 0$. The value of this limit is denoted as $1 - p$ and corresponds to the proportion of cured individuals in the population or cure rate [16]. Estimating the value of $1 - p$ is one of the main objectives of these models.

From their first appearance in 1949, various types of cure models have been proposed, which can be classified into two main groups: mixture cure models and promotion time cure models. The latter were designed as biological models for the analysis of relapse times in cancer studies [17]. They were formally proposed by [18], and they were initially used to model the tumour latency [19]. In such a case, it is assumed that after the first diagnosis and successful treatment, a number $N \geq 0$ of carcinogenic cells remain in the organism in a latent form, each one of them for a period of time T_k, until they finally develop a new tumour. Those individuals for whom $N \geq 1$ present at least one carcinogenic cell and are susceptible to relapse, whereas those with $N = 0$ are said to be cured and the latency time T is infinite [16].

Promotion time cure models assume that N follows a Poisson distribution with parameter $\theta > 0$, which is the average number of carcinogenic cells in the population. Assuming that the different T are i.i.d. with probability distribution $F(t)$, and independent from N, it can be demonstrated that the survival function of the population is defined by

$$S(t) = \exp[-\theta F(t)].$$

Mixture cure models are a sort of two-part models which were firstly introduced by [20]. These models study the response variables in two separate groups, which are identified by a binary variable, B, which is equal to 0 if the individual belongs to the cured group, and it is equal to 1 if the subject is susceptible to suffer the event of interest. Therefore, B is the indicator variable for the susceptibility, and it is only partially observed since it is not possible to distinguish between those susceptible observations that are censored and those observations of cured individuals.

In the context of survival analysis with a fraction of cured individuals, mixture models define the survival function of the population as:

$$S_{pop}(t|\mathbf{x{,}z}) = 1 - p(\mathbf{x}) + p(\mathbf{x})S_u(t|\mathbf{z}). \tag{5}$$

In (5), $\mathbf{X}$ and $\mathbf{Z}$ are two sets of covariates which might be equal or not, and $p(\mathbf{x}) = P(B = 1|\mathbf{X} = \mathbf{x})$ represents the probability of being susceptible given the value of the covariates, $\mathbf{X}$, and it is known as the incidence of the model. On the other hand, $S_u(t|\mathbf{z}) = P(T > t|\mathbf{Z} = \mathbf{z}, B = 1)$ is the survival function for the susceptible group, conditioned to the set of covariates, $\mathbf{Z}$, and it is known as the latency of the model [16]. The model formulation provides that the cure rate, $1 - p(\mathbf{x})$, depends only on $\mathbf{X}$, whereas the survival function of the susceptible group depends only on $\mathbf{Z}$. The fact of having these two groups of covariates separated for cured and uncured individuals allows us to consider external factors to have different influence in both groups of patients. This is the main advantage of mixture cure models, the methodology in which this paper is focused. Depending on the assumptions established for the latency and the incidence of the model, there are parametric, semiparametric and nonparametric approaches of cure models.

Parametric models Parametric mixture cure models were the first mixture cure models developed, and can be considered the basis of any further research in this field during the past 50 years. They were introduced by [20] to study a mouth cancer cohort who had been treated with a certain therapy, with the intention to model the relapse of these patients. [20] considered the cure rate as constant, and the survival function of the susceptible individuals was modeled according to a lognormal distribution with independence to any external covariate. [21] studied deeply the approach from [20], considering an exponential model for the latency, also with independence of covariates.

More than 30 years after the original proposal, the first cure models considering the influence of covariates were developed. [22] proposed a new model where the latency followed a Weibull distribution, dependent on a set the covariates, $\mathbf{Z}$, and modelling the incidence with a logistic function.

These parametric models show limited flexibility, due to the strict assumptions with respect to the latency and the incidence distribution. A generalization of the latter models, but still maintaining the parametric behavior, are the models based on the accelerated failure time (AFT). They assume the presence of covariates where their effects are fixed and multiplicative by the accelerated factor on the time scale [23]:

$$\log(T^*) = \beta_0 + \beta\mathbf{Z} + \sigma\varepsilon. \tag{6}$$

In (6), T^* is the survival time of the susceptible individuals, and σ is a scaling positive parameter. These models consider an error term (ε), whose density function is previously defined. AFT models were firstly proposed by [24] and later developed by [25]. Note that the aforementioned [20, 21] and [22] models can be derived from them, by giving specific values to the model parameters.

In all the cases, the estimation of the model parameters of parametric mixture models is performed using the maximum likelihood criterion, which is derived from classical survival models. The likelihood function for these models is defined as the product of two contributions: on the one hand, the censored observations and, on the other hand, the uncensored observations. It is not possible to distinguish between cured and uncured individuals in the censored part of the sample.

Semiparametric models Semiparametric models arise from the necessity to improve the flexibility of the aforementioned approaches in order to extract information from the sample to a greater extent. They are said to be semiparametric since this flexibility is usually assigned to the latency, whereas the incidence is still modeled using parametric methods - usually, assuming a logistic regression for $p(\mathbf{x})$. Depending on the assumptions made on the survival function of the susceptible group, we may find several types of semiparametric cure models:

- Proportional hazards (PH) cure models. These models are based on the regression model, which is applied to general survival studies in order to model the risks that may affect a population. Therefore, these models are based on the hazard function, which is directly related to the survival function. In PH models, the hazard function of an individual is given by the product between the baseline function and a non-negative function of the covariates:

  $$h(t|\mathbf{z}) = h_0(t)c(\boldsymbol{\beta}'\mathbf{z}).$$

  The function $c(\cdot)$ is known as the *link function* and is frequently chosen as the exponential function [2]. Besides, $h_0(t)$ is the baseline hazard function and may take any possible shape, being parametric or nonparametric. Given the relationship between the hazard and the survival functions, the model can be expressed in terms of this latter function:

  $$S(t|\mathbf{z}) = S_0(t)^{c(\boldsymbol{\beta}'\mathbf{z})},$$

  where S_0 represents the survival baseline, defined as $S_0(t) = P(T > t|\mathbf{Z} = 0, B = 1)$. These models were introduced by [26], who adapted the parameter estimation methods for classical survival analysis to the presence of a cure fraction. Due to the additional condition $B = 1$ in the definition of S_0, the traditional estimation methods cannot be applied, and thus additional tools were developed, based on expectation-maximization (EM) methods [16].
- AFT models. These models consist of a semiparametric adaptation of the aforementioned AFT models for the latency. These were introduced by [27], allowing the error term to present any survival functions without restrictions. Similarly as in the PH models, the incidence is modeled using a logistic function. The parameter estimation is performed using the maximum likelihood criterion, via EM methods.
- Flexible models. The aforementioned semiparametric approaches considered a logistic regression for incidence estimation. Nonetheless, [28] proposed an enhancement in the model flexibility by introducing

new functions to infer the cure rate, such as the probit or log-log distributions. They also considered the EM algorithm and the maximum likelihood criterion to estimate the parameters.

This flexibility, however, might not be enough to gather all the information within the study since incidence estimation is still parametric. Although parametric methods show some important advantages, such as their ease of interpretation or the simplicity of the parameter estimation, some authors have proposed semiparametric approaches to estimate the cure rate. Some of these semiparametric approaches are based on splines [29] or on single-index structures [30]. The flexibility can be further improved if completely nonparametric forms are introduced for the latency, being independent of any external covariate [31] or even taking these external factors into account [15], while the incidence is still modeled using parametric formulations.

Nonparametric models In the aforementioned cases, at least one of the model components, namely the latency or the incidence, was defined by a (semi-)parametric formulation. However, a completely nonparametric approach for both elements can be also considered, achieving thus the maximum flexibility of the models. The first nonparametric models were proposed by [32], but the authors did not consider the influence of covariates in the latency and incidence. This issue was partially solved by [33], who considered discrete variables.

The main progress in this field was carried out by [10], who proposed a completely nonparametric model, based on the nonparametric estimator for the cure rate developed by [34]. This estimator is also based on the Beran estimator for the survival function (see (3)). Regarding the selection of the smoothing parameters in this nonparametric context, [35] introduce a bootstrap bandwidth selection method for both the latency and the cure rate estimations. Furthermore, [36] proposed a nonparametric covariate hypothesis test for the incidence in mixture cure models, which can be applied to continuous, discrete and qualitative variables. This test allows for the identification of those variables that play a significant role on the cure rate. This nonparametric model does not assume any previous restriction, and therefore, it can be completely adjusted to the data.

3 Mixture cure models applied to COVID data

3.1 Dataset

A practical study has been performed using a COVID-19 database to illustrate the application of cure models. COVID-19 databases are broadly studied in the present time, as reviewed in [37], to develop a variety of mathematical models on the disease features [38]. The data was extracted from the *Servicio Galego de Saúde* and provided by the *Dirección Xeral de Saúde Pública* (Galicia, NW Spain). The dataset consists of 4307 COVID-19 patients who tested positive by PCR, being thus a representative cohort which has also been used to model several features related to COVID-19 disease, such as the disease severity [39] and the hospital and intensive care unit (ICU) length-of-stay [40]. The available variables for each individual are:

- ID number of the patient, unique and anonymous
- The age of the individual at diagnosis
- The sex (male/female) of the patient
- The date when the PCR test was first performed
- The hospital admission date, in case it was necessary due to the severity of the symptoms

The variable of interest is defined as "time from diagnosis to hospital admission". Hospital admission and bed occupation is an important issue that needs to be addressed in order to cope with the current situation, and thus it has been the target of a large number of investigations [41–48]. Since this is a time-to-event variable, survival methods can be used, as [49] did in their investigations concerning this variable for a Catalonian cohort [49]. Furthermore, there is a large fraction of individuals in the dataset (around 90%) who had not been admitted into hospital at the end of the follow-up period. Part of these individuals might be censored observations and thus, admission would occur after the end of the study. However, there will also be a number of individuals which do not require hospitalization during the illness, being those cured observations.

The database analysis by means of survival and cure models was performed using R software [50, 51].

3.2 Preliminary analysis

From the 4307 patients of the sample, 2615 (60.72%) are female. The average age is 57.1 years. The average age for men is 56.3 years and for women is 57.6 years. The distribution of the sample in terms of age and sex is represented in Fig. 1.

Data registration started on March, 6th, 2020, and the first hospitalization occurred a day after. The last diagnosis observation is placed on April, 2nd, whereas the last observed hospitalization took place on April, 3rd. The minimum survival time, defined as the elapsed time from diagnosis until hospital admission is 1 day, being 13 days the maximum uncensored observation of the time-to-event variable.

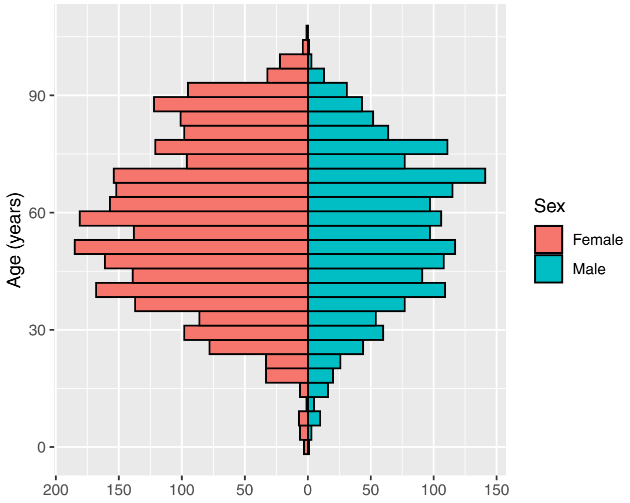

Fig. 1 Demographic distribution of the sample in terms of age and sex

A first insight of the study variable behavior can be performed using the KM estimator for the survival curve (Fig. 2). By using this approach, it is possible to assess some of the main features of the dataset. The points marked with the + symbol correspond to censored observations, which are distributed all over the time line. The jumps represent the time points where the failure has occurred for the uncensored observations. The form of the curve is due to the characteristics of the individuals in the database: the diagnosis and hospitalization times were registered in a short date format, considering only the day of diagnosis. Thus, the survival times are discrete observations and, considering the relatively short extension of the study, there exist large leaps that finally lead to the observed curve shape. Furthermore, we can also observe a *plateau* in the curve, which extends from the last uncensored observation on.

In the same way, a preliminary analysis for the covariate influence (age and sex of the patient) can also be performed by using the Beran estimator presented in (3). Specifically, the survival curves conditional on the age and sex were calculated, and the shape of these curves was observed for several values of these covariates. In Fig. 3 a comparison of the curves obtained with different covariate values is presented, when the "sex" variable is equal to 0 (male) or 1 (female), or when the covariate "age" is equal to the 1st or 3rd sample quantiles, namely 42 and 72 years. Apparently, the age covariate has an increased influence in the survival when compared to the covariate "sex", since the conditional survival functions are more far apart.

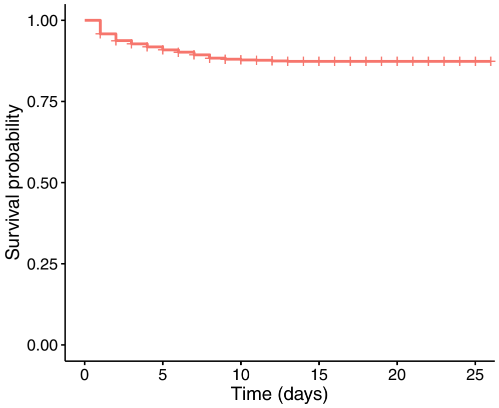

Fig. 2 Survival function estimation using the KM method for time to hospital admission

In order to obtain further information about the behavior of the study variable, it is necessary to apply the aforementioned cure models. Therefore, in this study we will apply semiparametric approaches, implemented in the *smcure* R package by [50], and nonparametric approaches, using functions from the *npcure* R package by [51].

3.3 Semiparametric cure models

The *smcure* package contains the tools needed to apply mixture cure models to a given dataset in a semiparametric context [50]. It considers Cox PH-derived models and AFT-derived models, as those presented in Section 2. The implementation of the *smcure* function allows us to explicitly select the model, as well as to select the type of parametric regression for the incidence between the three following distributions: logit, probit and complementary-loglog (cloglog). The parameter adjustment is performed using the EM algorithm under the maximum likelihood criterion.

A total of 6 different semiparametric mixture cure models were defined by combining the two available models for the latency (PH and AFT models) with the three possibilities for the incidence regression (logit, probit and cloglog). In all the models, both the sex and the age of the patient were considered as covariates for both the latency and the incidence. The significance of these covariates obtained for each of the models is presented in Tables 1 and 2. The covariates could not be considered as significative in any of the cases, and thus we cannot claim that the age or the sex of an individual affects the probability of needing hospitalization or the time since diagnosis until hospital admission.

The AFT and Cox PH models were compared by setting the sex and age covariates to some representative values in the sample. In this case, the survival curves were estimated considering a female individual since females represent the majority of the sample, 57 years old, which is the median age within the sample. The survival curves for this representative individual estimated with both latency models (AFT and PH), applying a probit regression for the incidence, are presented in Fig. 4. As we can observe,

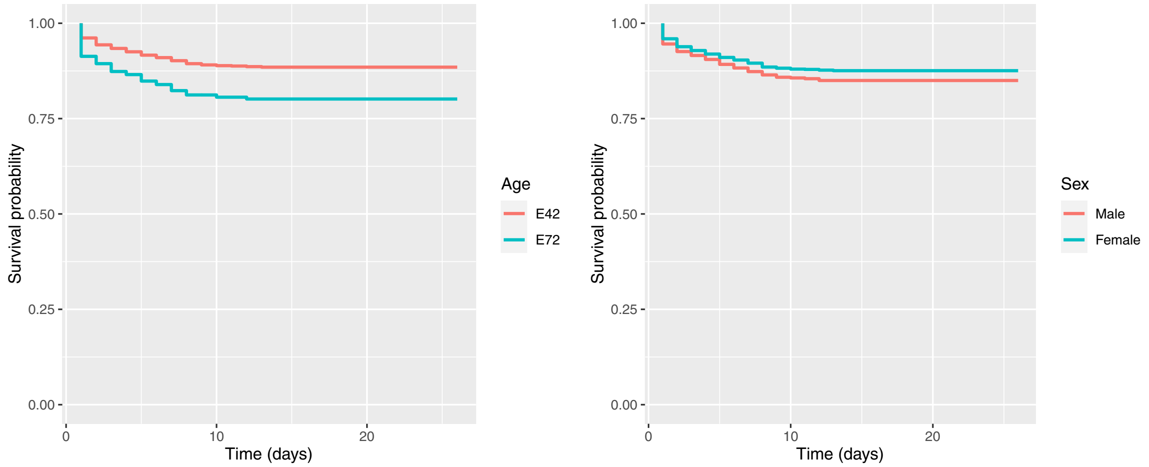

Fig. 3 Estimated survival curves obtained by the Beran method, which considers the survival conditioned to a set of covariates. On the left, the survival curve estimated for two individuals aged 42 and 72 years, aiming to illustrate the influence of this variable. On the right, the estimated survival curves of women and men, which allow us to observe the influence of the sex of the patient on the study

there is not a significant difference between both models. This can be due to the intrinsic limitations in the dataset used for our analyses. In this particular case, we can anticipate, on the one hand, the presence of a low fraction of susceptible individuals, since about 90% of the observations are censored. On the other hand, the time-to-event variable is discretized, since the starting point considered is the day when the diagnosis took place, ignoring the exact moment within it. Even though there are some mixture cure models that specifically consider discrete time variables [52], they have not been implemented in R yet.

In order to tackle with the aforementioned limitations, we simulated the behavior of a continuous-time variable using as a basis the original dataset. For this, we perturbed the data with random variables derived from a uniform distribution $U(-1, 1)$. As a result, a value between 0 and 1 is randomly subtracted or added to the original value of the time variable, and therefore the lifetimes will no longer be represented by integers.

After applying this procedure and fitting again the previously presented models, it was found that the covariate "sex" presented a significant influence on the latency of the AFT models, with a p-value 0.015. This fact is also evident when representing the estimated survival curves for women and men, where we can notice that the survival of women is higher when compared to men. Thus, we can conclude that the period from the diagnosis until the hospital admission for those patients with severe symptomatology will be longer for women. The variable "sex", however, was not found significant for the incidence of the model, concluding that with this approach men and women show equal probabilities of needing hospitalization. This reinforces the ideas by [53] and [54]. However, recent researches on COVID-19 cohorts have shown that male patients exhibit a worse prognosis compared to women, which may lead to an early need for hospitalization, as shown by these results [54].

Furthermore, this analysis allows proper comparison between AFT and PH models. Besides the different results obtained for each of the models, in Fig. 5 (right), we can see that for a 57 years old female patient, a slightly different behavior between models is appreciated. This difference reinforces the importance of an accurate selection of the

Table 1 Significance of the covariates "sex" and "age" on the latency of the proposed semiparametric models

Latency model	Variable	Coefficient estimation	p-value
Proportional hazards Cox model	age	$1.26 \cdot 10^{-3}$	0.641
Proportional hazards Cox model	sex	$2.63 \cdot 10^{-2}$	0.744
Accelerated failure time	age	0.00	1.00
Accelerated failure time	sex	0.00	1.00

Table 2 Significance of the covariates "sex" and "age" on the incidence for the three approaches considered

Incidence model	Variable	Coefficient estimation	p-value
logit	age	$4.02 \cdot 10^{-8}$	0.186
logit	sex	$-1.66 \cdot 10^{-7}$	0.934
probit	age	$-2.95 \cdot 10^{-11}$	0.999
probit	sex	$4.51 \cdot 10^{-10}$	0.999
cloglog	age	$-3.04 \cdot 10^{-11}$	0.932
cloglog	sex	$2.42 \cdot 10^{-9}$	0.908

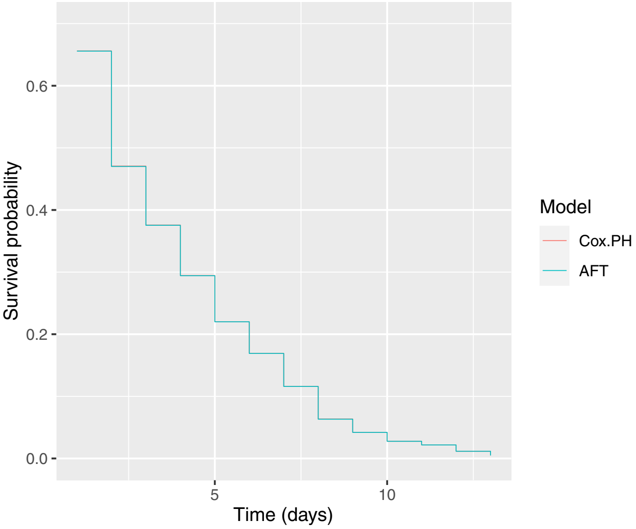

Fig. 4 Comparison between the latency estimated with both AFT and PH models for the same individual. The survival probability values are so close between models that is nearly impossible to distinguish both curves

model characteristics in order to capture the information within the sample.

3.4 Nonparametric cure models

The *npcure* package by [51] can be used to apply nonparametric estimations for mixture cure models. It includes the methods developed by [10, 35] and [36], which have been previously presented in Section 2.

Besides the appropriate tools needed to study the latency and the incidence of the population for a certain event of interest, this package includes some covariate hypothesis tests that can be used to confirm the inclusion of a certain covariate in the incidence in a mixture cure model.

Furthermore, the test by [32] is also implemented, which allows us to check if there is a cured fraction of individuals in the sample, and thus the suitability of cure models. Accepting or rejecting the null hypothesis determines the absence or presence of a pleateau in the survival curve, respectively. This plateau corresponds exclusively to the observations of cured individuals. The Maller-Zhou test was applied to the dataset, and a p-value equal to 0 was found. Therefore, we can reject the null hypothesis and thus confirm that there exists a cured fraction of individuals within the sample.

Another hypothesis test that can be used to study the significance of covariates for the cure probability is also included in the *npcure* package. This test is based on [55], and it was extended by [36], making it possible to determine whether the cure probability, is dependent of a given covariate X or not:

$$\begin{cases} H_0 : \text{cure probability} = 1 - p \\ H_1 : \text{cure probability} = 1 - p(x) \end{cases}$$

These hypotheses were tested for the COVID-19 database, and the results show that, with a significance level $\alpha = 0.01$, both sex and age influence significantly on the cure probability of the population, with $p-$values equal to 0.002 and 0, respectively. Thereby, we justify the inclusion of both covariates in our analyses.

In the *npcure* package, the estimation of the latency is implemented following the method described in [35], which is based on the Beran estimator for the calculation of the survival curve of the susceptible individuals, conditionally on a certain set of covariates. For each one of the covariates, namely age, sex, or a combination of both, a different curve is obtained. Since this estimation uses a kernel smoothing method, selection of the smoothing parameter is performed. The bootstrap method is used to mimic the minimization of the Mean Integrated Squared Error (MISE) criterion. The value of the bootstrap MISE is approximated using Monte Carlo, based on 100 bootstrap resamples.

In order to analyze the effect of the age on the latency, we estimated the survival curve when this covariate is equal to 20, 50 and 80 years, using the previously computed bootstrap smoothing parameter selector. The result is presented in the left part of Fig. 6. As it was expected, among the COVID-19 patients who needed hospitalization, those in their early stages of life tend to need it later after their diagnosis when compared with the elder population. The probability of the need for hospital admission at the beginning of the disease increases in the case of older people, and it is at that point in the course of the disease when the differences are more evident. This is consistent with the results of the epidemiological research that has been carried out in the last months, which claims that age is a clear risk factor for COVID-19 bad prognosis. Recently, it has been empirically observed that patients aged over 65 are prone to the need for ICU admission or respiratory support, at the same time they present a decreased lymphocyte count compared to young individuals [56].

As for sex, even though the covariate hypothesis test showed that this factor implied a significant influence on the cure rate (that is, no need of hospital admission), the estimated survival function of male and female were practically equivalent, as it can be observed in the right part of Fig. 6.

The covariate age also showed a significant influence on the incidence of the model, as it was anticipated by the covariate test (Fig. 7). In order to observe the changes in the cure probability when increasing the age of the patient, its value was estimated with the nonparametric method proposed by [10] for an age interval between 20

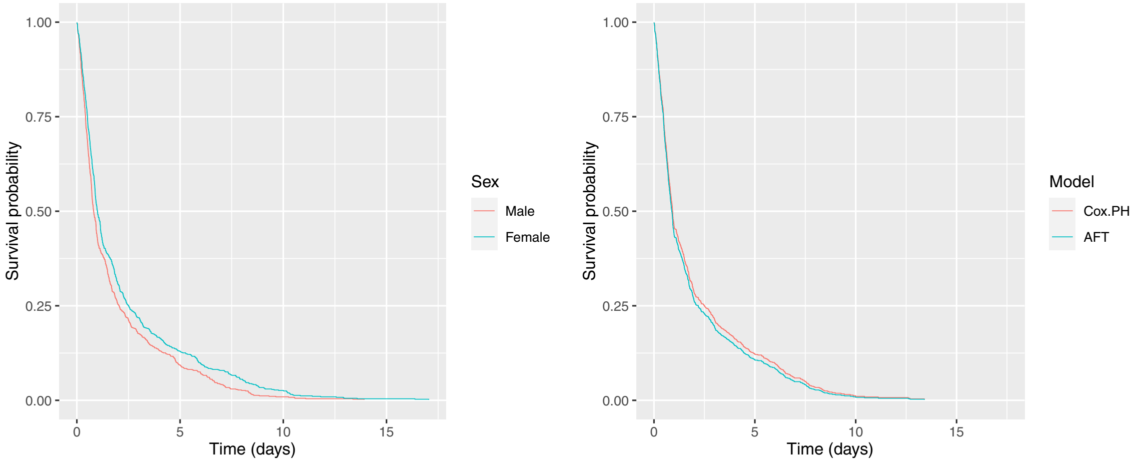

Fig. 5 Results obtained with the semiparametric models when working with continuous-time variables, which were calculated using the probit model for the incidence. On the left, the survival curves of susceptible patients are estimated separately for women and men using the AFT model. On the right, the survival curves are estimated for a 57 year-old female individual considering the two available semiparametric approaches

and 90 years old. The cure probability showed a decreasing tendency when increasing the age of the patient. This is also consistent with the previous results on COVID-19 literature, since it has been claimed that there is a clear increase in the hospitalization rates with advanced ages, being elder adults those who needed more frequently hospital care [57].

The covariate sex has a small influence in the cure probability, even though the covariate test considered a significant effect of this factor. This probability was found to be 0.870 for men and 0.877 for women. Sex, thus, apparently causes a difference of only 1% in the probability of hospitalization after a COVID-19 positive diagnosis. The difference between male and female patients, however, appeared to be significant when these probabilities were calculated in the interval of ages between 20 and 90 years (Fig. 7). It is important to highlight that males need, on average, hospital care more frequently than women, and this need clearly increases when considering elder patients. Once again, these results fall into line with the literature [54]: men tend to suffer a more severe symptomatology and thus, to need hospital admission more frequently than women.

In the case of nonparametric models, we also performed the data perturbation in order to obtain continuous-time data, following the same approach as with semiparametric models. The main difference when compared with the discrete data models was the loss of signification for sex in terms of cure probability (p-value $=$ 0.058), although the $p-$value obtained for this test is close to the significance level if we consider $\alpha = 0.05$. The age

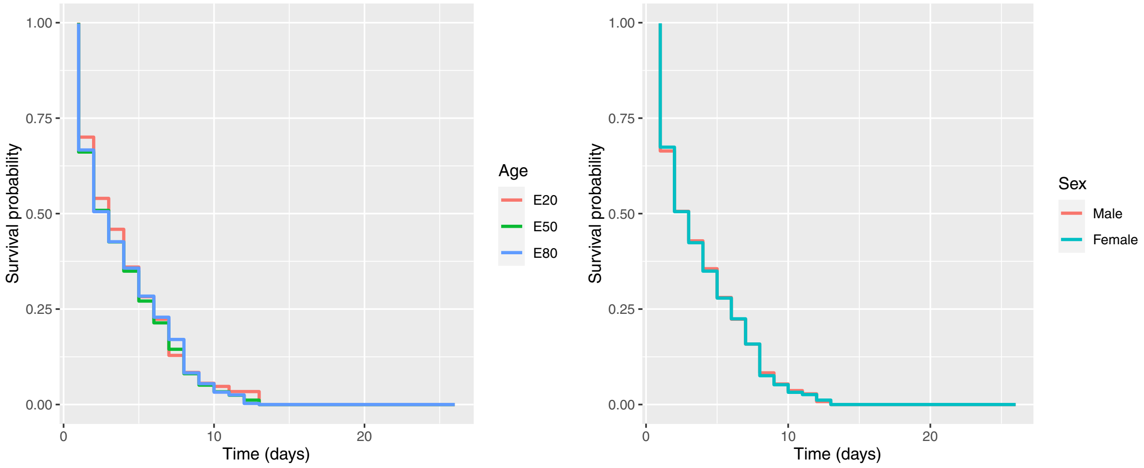

Fig. 6 Estimated survival function for the time until hospital admission of the susceptible patients using the nonparametric model, conditional on the covariates. On the left, model latency calculated for different values of the covariate "age of the patient". On the right, latency of the model conditional on the sex of the patient

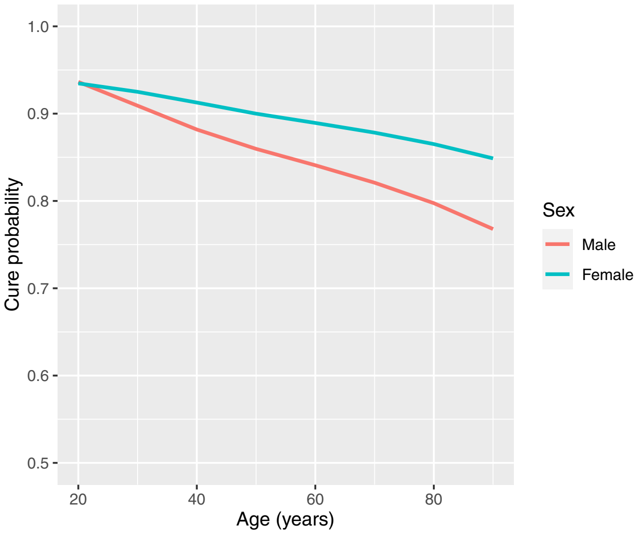

Fig. 7 Estimation of the cure probability (no need of hospital care) of COVID-19 patients with respect to the age, which shows a clear decreasing trend for both men and women

is still found significant for such a probability, and the estimated incidence is equivalent to the one obtained with the discretized times (Fig. 8, right side).

With regards to the latency, we obtained a smooth estimation of the survival curve when working with a continuous-time variable, which is more likely to be an accurate approximation of the real function. Furthermore, the differences between curves are easily observed when implementing the data perturbation. In the left side of Fig. 8, it can be noticed that the lifetime of young patients (that is, the elapsed time until their hospitalization), increases in the first stages of the disease, and then decreases until it is on the same level as the lifetime of the elder. This can be due to the current COVID-19 protocols: those patients over a certain age are directly admitted to the hospital when any minimal medical complication arises, whereas young individuals will only require hospital care if the symptoms are severe or if they last for a long time period.

4 Discussion

In our research, it has been claimed that different cure models lead to a variety of results and reach different conclusions, even though all of them belong to the same category of mixture cure models. When working with discrete lifetimes, nonparametric models apparently fit better to the actual situation, since there exist studies sustaining the influence of sex and age on COVID-19 prognosis. When using semiparametric models, we concluded that the set of covariates considered for this study does not yield any significant effect neither on their latency nor on their incidence, independently of the configuration selected for the model. Furthermore, it has been previously stated that these models might fail to reach convergence in the likelihood maximization, and thus resulting in biased estimators [58]. Due to the intrinsic features of our data, showing some limitations such as the discretization of the times, the high rate of censored data and a foreseeable high cure probability, it is possible that semiparametric models are not a good choice for analyzing this data. Nonetheless, they have been previously used for other COVID-19 studies, specifically semiparametric Cox PH models, which also considered the influence of the age on the survival [59].

In order to tackle with some of the limitations, we analyzed the behavior of the models when working with continuous data. In this case, it was indeed possible to

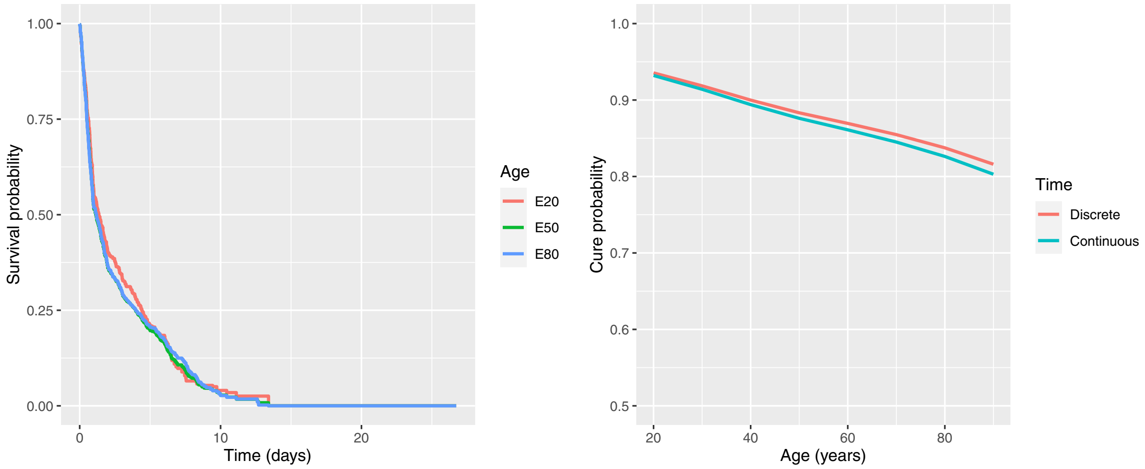

Fig. 8 Features of the nonparametric model when working with continuous data. On the left, comparison between survival curves of the patients with respect to their ages (20, 50 and 80). On the right, the trend of the cure probabilities when calculated with discrete and continuous times is compared

assess the effect of the sex of the patient on the latency. Therefore, the importance of working with exhaustive, high quality data in order to obtain feasible conclusions was reinforced.

Furthermore, nonparametric models were able to determine the effect of the study covariates on the time to hospital admission variable, even when it was discretized. With this approach, we studied the influence of the age on the survival of the population, and the results were consistent with the epidemiological data available in the literature. As far as we know, there are not previous publications considering completely nonparametric models to estimate the cure on COVID-19 cohorts, and this, together with the models developed by [40] on the length-of-stay prediction, are pioneer on the field.

Cure models are not the only tool that can be used to extract information about COVID-19 patients. Indeed, most of the aspects concerning this disease have been tackled using different sorts of models and algorithms, including clinical de-identification of COVID-19 datasets [60] or forecasting of the number of future patients using ARIMA models [61].

There also exist several studies aiming to predict the hospital and ICU admission based on different covariates such as age, gender and medical conditions of the patients, using classification algorithms such as Support Vector Machines and Random Forest [62, 63]. This leads to conclusions that can be completed with the results of this work, since this will give information of when that hospital admission will take place. However, machine learning-based approaches are not the best option when working with censored data, although they are specially helpful when handling high-dimensional clinical data [64]. Note that, in a context with censored data, it is not possible to apply directly machine learning classical models since they do not account for censored observations. Therefore, machine learning techniques should be adapted so that they also consider individuals who do not experience the event of interest. Thus, a few techniques have arisen in order to adapt this kind of predicting algorithms to the peculiarities of censored time-to-event observations, such as likelihood-based approaches [65] or the inverse probability of censoring [66]. Nonetheless, these methods are not completely developed and, as far as we know, theoretical results that prove their good behavior have not been presented yet. For future research, it will be interesting to study and propose a complete adaptation of machine learning techniques to the context of censored data. Therefore, both approaches will contribute in the analysis, leading to a more accurate result.

5 Conclusions

Over the last 60 years, there has been a remarkable progress in cure models and thus several implementations of these tools are available. In this work, we pointed out the differences among them, firstly from a theoretical perspective and later by their application to a real dataset. It has been observed that different model implementations reach a variety of conclusions related to the same dataset, which highlights the importance of using a suitable model when studying time-to-event data.

On the other hand, this paper emphasizes the importance of having high-quality datasets. It has been observed that working with discrete data leads to completely different results than considering continuous data. It is important to note that, besides choosing a suitable model for the data, using a representative and informative database is also an essential part in the analysis.

When working with proper variables and suitable models, it has been possible to analyze some key aspects of hospitalization in COVID-19 and its relationship with the variables sex and age of the patient. Therefore, it has been proved that cure models are helpful tools for this kind of studies.

Acknowledgements MPL activity was funded by the Science, Technology, and Innovation Plan of the Principality of Asturias (Spain) Ref: FC-GRUPIN-IDI/2018/000225, which is part-funded by the European Regional Development Fund (ERDF). ALC was sponsored by the BEATRIZ GALINDO JUNIOR Spanish Grant from MICINN (Ministerio de Ciencia, Innovación y Universidades) with reference BGP18/00154. RC and ALC acknowledge partial support by the MINECO grant MTM2017-82724-R, and by the Xunta de Galicia: Grupos de Referencia Competitiva ED431C-2020-14, Centro de Investigación del Sistema universitario de Galicia ED431G 2019/01, and Axencia Galega de Innovación (Ayudas proyectos de investigación COVID-19 presentados a la convocatoria del ISCIII IN845D 2020/26 - Programa Operativo FEDER Galicia 2014-2020), all of them through the ERDF.

References

1. Singh R, Mukhopadhyay K (2011) Survival analysis in clinical trials: Basics and must know areas. Perspect Clin Res 2:145–148. https://doi.org/10.4103/2229-3485.86872
2. Klein JP, Moeschberg ML (2003) Basic quantities and models. Survival analysis techniques for censored and truncated data. Springer, Nueva York. https://doi.org/10.2307/2281868
3. Kleinbaum DG, Klein M (2012) Introduction to survival analysis. Survival analysis a Self-Learning text. Springer, Nueva York. https://doi.org/10.1093/biomet/79.3.531
4. Prinja S, Gupta N, Verma R (2010) Censoring in clinical trials: Review of survival analysis techniques. Indian J Community Med 35:217–221. https://doi.org/10.4103/0970-0218.66859

5. Klein JP, Moeschberg ML (2003) Censoring and truncation. Survival analysis techniques for censored and truncated data. Springer, Nueva York. https://doi.org/10.2307/2281868
6. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481. https://doi.org/10.2307/2281868
7. Bewick V, Cheek L, Ball J (2004) Statistics review 12: Survival analysis. Crit Care 8:389–394. https://doi.org/10.1186/cc2955
8. Johansen S (1978) The product limit estimator as maximum likelihood estimator. Scand J Stat 5:195–199. https://doi.org/10.1080/03610928708829561
9. Beran R (1981) Nonparametric regression with randomly censored survival data. Tech. rep. University of California, Berkeley, Berkeley
10. López-Cheda A, Cao R, Jácome MA, Van Keilegom I (2017) Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. Comput Stat Data Anal 105:144–165. https://doi.org/10.1016/j.csda.2016.08.002
11. Dabrowska D (1989) Uniform consistency of the kernel conditional Kaplan-Meier estimate. Ann Stat 17:1157–1167. https://doi.org/10.1214/aos/1176347261
12. Iglesias-Pérez MC (2007) Selección de la ventana en estimación de la distribución condicional. In: Libro de Actas del XXX Congreso Nacional de Estadística e Investigación Operativa
13. Emmert-Streib F, Dehmer M (2019) Introduction to survival analysis in practice. Mach Learn Know Extr 1:1013–1038. https://doi.org/10.3390/make1030058
14. Ciuca V, Matei M (2009) Survival analysis for the unemployment duration. Proc 5th WSEAS Int Conf Econ Manag Transform 1:354–359
15. Patilea V, Van Keilegom I (2020) A general approach for cure models in survival analysis. Ann Stat 48:2323–2346
16. Amico M, Van Keilegom I (2018) Cure models in survival analysis. Annual Rev Stat Appl 5:311–342. https://doi.org/10.1146/annurev-statistics-031017-100101
17. Lambert P, Bremhorst V (2019) Estimation and identification issues in the promotion time cure model when the same covariates influence long- and short-term survival. Biom J 61:275–289. https://doi.org/10.1002/bimj.201700250
18. Yakovlev AY, Tsodikov A (1996) Stochastic models of tumor latency and their biostatistical applications. World Scientific Pub Co Inc, Singapore
19. Yakovlev AY, Asselain B, Bardou VJ, Fourquet A, Hoang T, Rochefediere A, Tsodikov AD (1993) A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. Biometrie et Analyse de Donnees Spatio-Temporelles 12:66–82
20. Boag JW (1949) Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J R Stat Soc Ser B - Stat Methodol 11:15–53. https://doi.org/10.2307/2983694
21. Berkson J, Gage RP (1952) Survival curve for cancer patients following treatment. J Am Stat Assoc 47:501–515. https://doi.org/10.2307/2281318
22. Farewell VT (1982) The use of mixture models for the analysis of survival data with long-term survivors. Biometrics 38:1041–1046. https://doi.org/10.2307/2529885
23. Saikia R, Barman MP (2017) A review on accelerated failure time models. Int J Stat Syst 12:311–322
24. Yamaguchi K (1992) Accelerated failure-time regression model with a regression model of surviving fraction: An analysis of permanent employment in Japan. J Am Stat Assoc 87:284–292. https://doi.org/10.1080/01621459.1992.10475207
25. Peng Y, Dear KB, Denham JW (1998) A generalized F mixture model for cure rate estimation. Stat Med 17:813–830. https://doi.org/10.1002/(SICI)1097-0258(19980430)17
26. Kuk AYC, Chen CH (1992) A mixture model combining logistic regression with proportional hazards regression. Biometrika 79:531–541. https://doi.org/10.1093/biomet/79.3.531
27. Li C, Taylor JMG (2002) A semi-parametric accelerated failure time cure model. Stat Med 21:3235–3247. https://doi.org/10.1002/sim.1260
28. Lam KF, Fong DYT, Tang OY (2005) Estimating the proportion of cured patients in a censored sample. Stat Med 24:1865–1879. https://doi.org/10.1002/sim.2137
29. Wang L, Du P, Lian H (2012) Two-component mixture cure rate model with spline estimated nonparametric components. Biometrics 68:726–735. https://doi.org/10.1111/j.1541-0420.2011.01715.x
30. Amico M, Van Keilegom I, Legrand C (2019) The single-index/Cox mixture cure model. Biometrics 75:452–462. https://doi.org/10.1111/biom.12999
31. Taylor JMG (1995) Semi-parametric estimation in failure time mixture models. Biometrics 51:899–907. https://doi.org/10.2307/2532991
32. Maller RA, Zhou S (1992) Estimating the proportion of immunes in a censored sample. Biometrika 79:731–739. https://doi.org/10.1093/biomet/79.4.731
33. Laska EM, Meisner MJ (1992) Nonparametric estimation and testing in a cure model. Biometrics 48:1223–1234. https://doi.org/10.2307/2532714
34. Xu J, Peng Y (2014) Nonparametric cure rate estimation with covariates. Can J Stat 42:1–17. https://doi.org/10.1002/cjs.11197
35. López-Cheda A, Jácome MA, Cao R (2017) Nonparametric latency estimation for mixture cure models. Test 26:353–376. https://doi.org/10.1007/s11749-016-0515-1
36. López-Cheda A, Jácome MA, Van Keilegom I, Cao R (2020) Nonparametric covariate hypothesis tests for the cure rate in mixture cure models. Stat Med 39:2291–2307. https://doi.org/10.1002/sim.8530
37. Shuja J, Alanazi E, Alasmary W, Alashaikh A (2021) COVID-19 Open source data sets: A comprehensive survey. Appl Intelli 51:1296–1325. https://doi.org/10.1007/s10489-020-01862-6
38. Mohamadou Y, Halidou A, Kaper PT (2020) A review ofmathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19. Appl Intell 50:3913–3925. https://doi.org/10.1007/s10489-020-01770-9
39. Gude F, Fernández-Merino C, Ferreiro L, Lado-Baleato O, Espasadín-Domínguez J, Hervada X, Cadarso CM, Valdés L (To appear) Development and validation of a prognostic model based on comorbidities to predict covid-19 severity. Int Jr Epidemiology
40. López-Cheda A, Jácome MA, Cao R, de Salazar PM (2020) Estimating COVID-19 hospital demand using a non-parametric model: A case study in Galicia (Spain). Unpublished manuscript https://doi.org/10.1101/2020.09.04.20187963
41. Lapidus N, Zhou X, F Carrat BR, Zhao Y, Hejblub G (2020) Biased and unbiased estimation of the average length of stay in intensive care units in COVID-19 pandemic. Unpublished manuscript https://doi.org/10.1101/2020.04.21.20073916
42. Li R, Rivers C, Tan Q, Murray MB, Toner E (2020) Estimated demand for US Hospital Inpatient and Intesive Care Unit beds for patients with COVID-19 based on comparisons with Wuhan and Guangzhou, China. JAMA Netw Open 3:(e208297) https://doi.org/10.1001/jamanetworkopen.2020.8297
43. Moghadas SM, Shoukat A, Fitzppatrick MC, Wells CR, Sah P, Pandey A, Sachs JD, Wang Z, Meyers LA, Singer BH, Galvani AP (2020) Projecting hospital utilization during the COVID-19 outbreaks in the United States. PNAS 117:9122–9126. https://doi.org/10.1073/pnas.2004064117/-/DCSupplemental

44. Qi X, Jiang Z, Yu Q, Shao C, Zang H, Yue H, Ma B, Wang Y, Liu C, Meng X, Huand S, Wang J, Xu D, Lei J, Xie G, Huang H, Yand J, Ji J, Pan H, Zhou S, Ju S (2020) Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study. Unpublished manuscript https://doi.org/10.1101/2020.02.29.20029603
45. Rees EM, Nighttingale ES, Jafaria Y, Waterlow NR, Clifford S, Pearson CAB, Group CW, Jombart T, Procter SR, Knight GM (2020) COVID-19 length of hospital stay: a systematic review and data synthesis. Unpublished manuscript https://doi.org/10.1101/2020.04.30.20084780
46. Thai PQ, Toan DTT, Son DT, Van HTH, Minh LN, Hund LX, Toan NV, Hoat LN, Luong DH, Khoa NT, Huong LT (2020) Factors associated with the duration of hospitalisation among COVID-19 patients in Vietnam: A survival analysis. Epidemiol Infect 348:1–7. https://doi.org/10.1017/S0950268820001259
47. Wang Z, Ji JS, Liu Y, Liu R, Zha Y, Chang X, Zhang L, Liu Q, Zhang Y, Dong T, Xu X, Zhou L, He J, Deng Y, Zhong B, Wu X (2020) Survival analysis of hospital length of stay of novel coronavirus (COVID-19) pneumonia patients in Sichuan, China. Unpublished manuscript https://doi.org/10.1101/2020.04.07.20057299
48. Wood RM, McWilliams CJ, Thomas MJ, Bourdeaux CP, Vasilakis C (2020) COVID-19 scenario modelling for the mitigation of capacity-dependent deaths in intensive cares. Health Care Management Science https://doi.org/10.1007/s10729-020-09511-7
49. Prieto-Alhambra D, Balló E, Coma E, Mora N, Aragón M, Prats-Uribe A, Fina F, Benítez M, Guiriguet C, Fábregas M, Medina-Peralta M, Duarte-Salles T (2020) Hospitalization and 30-day fatality in 121,263 COVID-19 outpatient cases. Unpublished manuscript https://doi.org/10.1101/2020.04.07.20057299
50. Cai C, Zou Y, Peng Y, Zhang J (2012) smcure: Fit Semiparametric Mixture Cure Models, R package version 2.0. http://CRAN.R-project.org/package=smcure
51. López-de-Ullibarri I, López-Cheda A, Jácome MA (2020) npcure: Nonparametric Estimation in Mixture Cure Models. https://CRAN.R-project.org/package=npcure, R package version 0.1-5
52. Zhao X, Zhou X (2008) Discrete-time survival models with long-term survivors. Stat Med 27:1261–1281. https://doi.org/10.1002/sim.3018
53. Gebhard C, Regitz-Zagrosek V, Neuhauser HK, Morgan R, Klein SL (2020) Impact of sex and gender on COVID-19 outcomes in Europe. Biol Sex Differ 11 https://doi.org/10.1186/s13293-020-00304-9
54. Jian-Min J, Peng B, Wei H, Fei W, Xiao-Fang L, De-Min H, Shi L, Jin-Kui Y (2020) Gender differences in patients with COVID-19: Focus on severity and mortality. Front Public Health 8:152. https://doi.org/10.3389/fpubh.2020.00152
55. Delgado MA, González-Manteiga W (2001) Significance testing in nonparametric regression based on the bootstrap. Ann Stat 29:1469–1507. https://doi.org/10.1214/aos/1013203462
56. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW (2020) Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City Area. JAMA 323:2052–2059. https://doi.org/10.1001/jama.2020.6775
57. Garg S, Kim L, Whitaker M (2020) Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019 — COVID-NET, 14 States, March 1–30. MMWR Morb Mortal Wkly Rep 69:458–464. https://doi.org/10.15585/mmwr.mm6915e3
58. Lu W (2010) Efficient estimation for an accelerated failure time model with a cure fraction. Stat Sin 20:661–674. https://doi.org/10.1002/sim.1260
59. Sreedevi EP, Sankaran PG (2020) Statistical methods for estimating cure fraction of COVID-19 patients in India. medRxiv 2020053020117804 https://doi.org/10.1101/2020.05.30.20117804
60. Catelli R, Gargiulo F, Casola V, Pietro GD, Fujita H, Esposito M (2020) Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set. Appl Soft Comput 97 https://doi.org/10.1016/j.asoc.2020.106779
61. Hernandez-Matamoros A, Fujita H, Hayashi T, Perez-Meana H (2020) Forecasting of COVID19 per regions using ARIMA models and polynomial functions. Appl Soft Comput 96 https://doi.org/10.1016/j.asoc.2020.106610
62. Hernández-Pereira E, Fontenla-Romero O, Bolón-Canedo V, Cancela B, Guijarro-Berdiñas B, Alonso-Betanzos A (2020) Authomatic classification of hospitalization of COVID-19 patients using machine learning. Unpublished manuscript
63. Davila-Pena L, García-Jurado I, Casas-Méndez B (2020) Assessment of the influence of the features in a classification problem: an application to the classification of COVID-19 patients. Unpublished manuscript
64. Spooner A, Chen E, Sowmya A, P Sachdev NAK, Trollor J, Brodaty H (2020) A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Scientif Rep 10(20410):7226–7234. https://doi.org/10.1038/s41598-020-77220-w
65. Stajduhar I, Dalbelo-Basic B (2012) Uncensoring censored data for machine learning: A likelihood-based approach. Expert Syst Appl An Int J 39:7226–7234. https://doi.org/10.1016/j.eswa.2012.01.054
66. Vock DM, Wolfson J, Bandyopadhyay S, Adomavicius G, Johnson PE, Vazquez-Benitez G, O'Connor PJ (2016) Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. J Biomed Inform 61:119–131. https://doi.org/10.1016/j.jbi.2016.03.009

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

Maria Pedrosa-Laza is a novel reseacher who received her Bachelor in Biotechnology form the University of Oviedo, and afterwards she obtained her M.S. in Bioinformatics for Health Science at the University of A Coruña (UDC). She is also a researcher at University of Oviedo. Pedrosa-Laza is currently interested in the application of statistics, artificial intelligence and data analysis to the field of clinical and non-clinical data.

Ana López-Cheda is a distinguished researcher within the Beatriz Galindo Junior Programme in the University of A Coruña (UDC) since November 2019. López-Cheda obtained all her degrees at the UDC: the degree of Computer Engineering; the Master's degree in Statistical Techniques, and the PhD degree in the PhD Program of Statistics and Operational Research. López-Cheda's research work has been, mainly, associated with survival analysis and its application to medical databases. Specifically, her current interests include cure models, functional data, nonparametric statistics and survival analysis.

Ricardo Cao received his Bachelor, M.S. and PhD degrees in Mathematics from the University of Santiago de Compostela, Spain. He is full professor in Statistics and Operations Reseearch and the Editor-in-Chief of the Journal of Nonparametric Statistics. He is also the head of the committee of experts on Mathematics versus COVID-19 in Spain, created by the Spanish Committee on Mathematics (CEMat). His current interests include nonparametric statistics, bootstrap, survival analysis, functional data analysis, big data statistical analysis, dependent data, empirical likelihood, credit risk and statistical methods in genomics, neuroscience and weed science.