# Main Manuscript for

## Estimating COVID-19 hospital demand using a non-parametric model: a case study in Galicia (Spain)

**Authors**: Ana López-Cheda[1], María-Amalia Jácome*[1], Ricardo Cao[2], Pablo M De Salazar[3]

**Affiliations**
1. Universidade da Coruña, CITIC, MODES, A Coruña, Spain
2. Universidade da Coruña, CITIC, ITMATI, MODES, A Coruña, Spain
3. Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard School of Public Health, Boston, US

* Corresponding author: María-Amalia Jácome.
Faculty of Science, Rúa da Fraga 10, 15008 A Coruña (Spain)
**Email:** maria.amalia.jacome@udc.es.

**ORCIDs**: Ana López-Cheda (https://orcid.org/0000-0002-3618-3246)[1], María-Amalia Jácome (https://orcid.org/0000-0001-7000-9623)[1], Ricardo Cao (https://orcid.org/0000-0001-8304-687X)[2], Pablo M De Salazar (https://orcid.org/0000-0002-8096-2001)[3]

**Keywords**

COVID-19; ICU; nonparametric; mixture cure model; forecasting, length of stay

**Author Contributions**

All authors contributed to the study and model design. ALC, MAJ and RC implemented the models and performed statistical analysis. All authors interpreted results and contributed to writing of the manuscript.

**Abstract**

Understanding the demand for hospital beds for COVID-19 patients is key for decision-making and planning mitigation strategies, as overwhelming healthcare systems has critical consequences for disease mortality. However, accurately mapping the time-to-event of hospital outcomes, such as the length-of-stay in the ICU, requires understanding patient trajectories while adjusting for covariates and observation bias, such as incomplete data. Standard methods, like the Kaplan-Meier estimator, require prior assumptions that are untenable given current knowledge. Using real-time surveillance data from the first weeks of the COVID-19 epidemic in Galicia (Spain), we aimed to model the time-to-event and event probabilities of patients hospitalized, without parametric priors and adjusting for individual covariates. We applied a nonparametric Mixture Cure Model and compared its performance in estimating hospital ward/ICU lengths-of-stay to the performances of commonly used methods to estimate survival. We showed that the proposed model outperformed standard approaches, providing more accurate ICU and hospital ward length-of-stay estimates. Finally, we applied our model estimates to simulate COVID-19 hospital demand using a Monte Carlo algorithm. We provided evidence that adjusting for sex, generally overlooked in prediction models, together with age is key for accurately forecasting ICU occupancy, as well as discharge or death outcomes.

**Main Text**

**Introduction**

As of September 2020, SARS-CoV-2 transmission continues to increase in most countries worldwide [1], and in those countries where control has been achieved, resurgences are expected [2] before an effective vaccine is widely available. Within the main challenges of the pandemic, overwhelming healthcare systems has critical consequences on disease mortality [3]. Thus, understanding and predicting inpatient and critical-care demand remains one of the major components of outbreak monitoring for decision-making and contingency planning.

Predicting hospital demand entails estimating a patient's length-of-stay (LoS) in a hospital ward or in the ICU. Estimating the LoS from data is challenging as it requires investigating the patients' trajectories, and it must account for complexities in the processes and the availability of data. For example, some outcome data may be missing because the study ends before the patient leaves the hospital; this missing data is referred to as right censored data. The duration of hospitalization of COVID-19 patients has been studied using parametric models [4], semiparametric methods [5], and nonparametric estimators [3, 6].

Parametric and semiparametric approaches are often preferred due to their simplicity and ease of interpretation, but they require the LoS to conform to a predefined fixed model. Estimations based on non-validated assumptions can be significantly biased. Thus, nonparametric approaches, which do not require model assumptions, should be used when estimating COVID-19 LoS in the absence of solid knowledge.

The Kaplan-Meier (KM) estimator [8] is the simplest and most often used nonparametric estimator in medical survival analysis of time-to-event data. It assumes that all patients with missing outcomes would experience the event in the end. This assumption applies when analyzing the duration of hospitalization, that is, the total time in the institution of the hospital (which includes time in hospital ward and time in ICU), as all patients eventually leave the hospital. However, this assumption does not apply to a patient's LoS in the hospital ward until admission to the ICU or until death (that is, not all hospital patients experience admission to the ICU or death). Thus, the KM estimator should not be used to estimate those LoS, as it is wrongly specified. Alternatively, Mixture Cure Models (MCM) [9] account for the situations when it is known that a proportion of individuals will not experience the event being analyzed.

Here, we propose a nonparametric Mixture Cure Model (NP-MCM) for estimating the lengths-of-stay until specific final events that are not experienced by all the patients (Safari et al., 2020). We estimated -in a completely nonparametric way without any dependence on preliminary model assumptions- the following 5 lengths-of-stay: LoS in hospital ward until admission to ICU, LoS in hospital ward until discharge from hospital ward, LoS in hospital ward until death in hospital ward, LoS in ICU until discharge from ICU; and LoS in ICU until death in ICU. We also estimated the probability of each event. To illustrate how our model improves data fitting, we compared the NP-MCM to the standard KM estimator (which assumes that all the individuals will experience the final event) and to the empirical (E) estimator (which discards all observations which event is not observed) for a dataset of COVID-19 patients from the first weeks of the epidemic in Spain. We further simulated inpatient and critical care cumulative incidence during an outbreak, along with the final outcome (discharge or death), using the estimated values, and adjusting for age and sex. Our model shows the importance of these individual variables for predicting hospital demand during transmission.

## Materials and Methods

### *Data source*

The dataset used in this paper contains 10454 confirmed COVID-19 cases reported in Galicia, a region in Spain's northwest, from March 6 to May 7, 2020. Data was provided for analysis by the regional public health authority, Dirección Xeral de Saúde Pública [10]. The data included information on age and sex; the dates of COVID-19 diagnosis, admission to the hospital and/or ICU; and the patient's last known clinical status. A summary of the dataset can be found in the Appendix.

### *Model formulation*

Mixture cure models [9], a special case of cure models [11], explicitly model survival as a mixture of two types of patients: those who will experience the final outcome and those who will not (that is, they are "cured" and therefore will not experience the event). Note that here a "cured" individual is defined as being free of experiencing the event of interest and is not necessarily cured in medical terms. The goal of MCM is to estimate the probability of experiencing the event and the distribution of the time to the event. The model is formulated as follows.

Let us denote Y as the time to the event of interest (admission to ICU, death, or discharge), with survival function $S(t) = P(Y > t)$. Let $p = P(Y < \infty)$ be the probability that the event will happen, and $S_0(t) = P(Y > t \mid Y < \infty)$ be the survival function of the individuals experiencing the event. MCM write the survival function as $S(t) = (1 - p) + pS_0(t)$. Then the probability of the event, p, and the survival function of the time-to-event, $S_0(t)$, can be estimated using a proper estimator of the survival function, S(t), and the relations:

$$p = 1 - S(t) \; and \; S_0(t) = \frac{S(t) - (1-p)}{p} \qquad \text{(eq1)}$$

Under right censoring, the observations are not $\{y_i, i = 1,\ldots,n\}$ but $\{(t_i, d_i, x_i), i = 1,\ldots,n\}$ with $t_1 \leq t_2 \leq \ldots \leq t_n$, where $t_i$ is the observed time, $d_i$ the indicator of whether the final outcome has been observed, and $x_i$ the indicator of whether patient i is known not to experience the event (cured). Hence, patients can be classified into three groups: (a) the event is observed ($t_i = y_i$, $d_i = 1$, $x_i = 0$); (b) the final outcome is not observed and it is unknown if it would have happened eventually ($t_i < y_i$, $d_i = 0$, $x_i = 0$); and (c) the event is not observed because it is known that it will never happen ($t_i < y_i$, $d_i = 0$, $x_i = 1$). In classical survival analysis when it is assumed that all the patients will experience the final outcome, only groups (a) and (b) are considered.

When there is a group of patients known not to experience the event, the survival function S(t) can be estimated nonparametrically as follows (Safari et al, 2020):

$$\hat{S}(t) = \prod_{t_i<t}\left(1 - \frac{d_i}{n-i+1+\sum_{j=1}^{i} x_j}\right) \quad \text{(eq2)}$$

To note, this estimator reduces to the well-known KM estimator in a classical time-to-event analysis when the event happens for all patients.

The estimator of S(t) in (eq2) is computed with R software [2] and used to estimate the probability of the event, p, and the time-to-event survival function $S_0(t)$ using the relationships in (eq1) for the following 5 lengths-of-stay: a) LoS in hospital ward until admission to ICU; b) LoS in hospital ward until death in hospital ward; c) LoS in hospital ward until discharge; d) LoS in ICU until death in ICU; e) LoS in ICU until ICU discharge. Details on each LoS, along with an R script for the computation of the different estimators, can be found in the Appendix.

The NP-MCM survival estimator of $S_0(t)$ is compared to the KM estimator computed with two different datasets: (a) the complete set of observations, considering as simply right censored all the patients who did not experience the event, regardless if they might experience it in the future or not (complete KM), and (b) a reduced dataset, dismissing the patients who, it is known, will not ever experience the event (reduced KM). The empirical (E) estimator, which considers only patients whose final event is observed and disregards the right censored observations, has also been considered.

The NP-MCM estimator of the probability, p, of the event was computed using the estimator of S(t) in (eq2) and the relationships in (eq1). The empirical estimator of p, given by the ratio between the number of observed events and the total number of patients, was computed to motivate the proposed NP-MCM estimator of p.

As for the KM estimator, the NP-MCM estimator in (eq2) does not incorporate possible covariate effects, such as those of sex and age. The extension of the KM estimator to handle covariates is the generalized product-limit estimator [13] of the conditional survival function, S(t|x). When the final outcome is not experienced by all the patients but only a group of them, the incorporation of covariates for the estimation of $S_0(t|x)$ has been studied recently. Specifically, if no patients in the dataset can be distinctly identified as being free from the event, the estimator of $S_0(t|x)$ and the probability of the final outcome p(x) [14-16] are implemented in the R package npcure [17], which also performs significance tests for the cure probability. The extension of these methods to situations where some patients are clearly known not to experience the final outcome, as it happens for our COVID-19 data, has been recently addressed (Safari et al., 2020), where evidence of the superiority of the NP-MCM over the traditional methods is shown. These conditional estimators of S(t|x) and $S_0(t|x)$ can handle continuous covariates such as age, using the information from all the individuals to provide estimates of the survival function for one single value of the covariate, e.g., 40 years. Ignoring the effect of age and sex on these time estimates can produce important bias in the statistical analysis.

*COVID-19 outbreak simulation model*

We further simulated a COVID-19 outbreak based on the NP-MCM estimates of the 5 lengths-of-stay considered, with two different models: 1) the simplest possible where the distributions of times and probabilities of moving from one state (hospital ward, ICU) to another (hospital ward, ICU, death, discharge) do not depend on individual covariates; and 2) a more realistic one with the LoS and transition probabilities depending on the available covariates of age and sex.

The simulated outbreak consisted of N = 1000 infected individuals. For the i-th infected individual i= 1, …, N we simulated the sex $G_i$ (0 = male, 1 = female) and the age $A_i$ (years) using the real distributions of the reported COVID-19 cases in Galicia on May 7, 2020 (see **Table 2** in the Appendix for details in case counts). Let $H \in \{1, \ldots, N\}$ be the set of indices corresponding to

infected subjects admitted to the hospital. The trajectory of every hospitalized patient $i \in H$ is obtained by simulating the transitions between states (hospital ward, ICU, discharge, death) using the NP-MCM estimated probabilities, and the times in each state were simulated from the Weibull distributions that best fitted the NP-MCM estimates, both in the unconditional setting as in the case of conditioning on the age and sex of the patient (see Appendix for further details; **Figure 8** for the density estimations; **Figure 9** for the survival curves; **Table 3** gives a numerical summary of each simulated time, **Table 4** shows the Weibull parameters of all the time distributions). Using these estimated times and going through all the hospitalized patients, it is straightforward to compute the number of patients in every state as a function of time. The mean number of reported cases and the mean number of patients in a hospital ward, in the ICU, who have died, and who have been discharged can be approximated by a Monte Carlo simulation as a function of time.

## Results

Using a dataset of hospitalizations of COVID-19 patients in Galicia (Spain) during the first weeks of the outbreak, we first compared the NP-MCM estimates with estimates from the E estimator, and the KM estimator by (a) treating as right censored all the data from patients who did not experience the event, whether or not they might experience it in the future (hereafter referred to as complete KM), and by (b) dismissing the data from patients who we knew would never experience the event (hereafter referred to as reduced KM). When an event ("leave the hospital") happens for all patients, the LoS estimates from the NP-MCM and the KM estimators are exactly the same, while the E estimator underestimates the LoS (see **Figure 1** for the survival estimates, and **Figure 6** in the Appendix for the corresponding density estimations). The NP-MCM and KM estimators consider n = 2453 patients who have ever been hospitalized, and 2142 patients who experienced the event (that is, they left the hospital within the study's timeframe). The E estimator considers only n = 2142 patients who left the hospital (discharged or died), disregarding the information from the 311 patients still in the hospital. This biases the E estimate toward shorter LoS, as hospitalized patients with longer LoS cannot be included in the estimation.

**Figure 1.** Estimates of the survival function of LoS using NP-MCM (thick black line), KM with the complete dataset (thin grey line), KM with the reduced dataset (thin black line) and the empirical E estimator (red line) for all the COVID-19 hospitalized cases (n = 2453) in Galicia (Spain), when the LoS is the duration of hospitalization (top left), time in hospital ward until admission to ICU (top right), time in hospital ward until death in hospital ward (middle left), time in hospital ward until discharge (middle right), time in ICU until death in ICU (bottom left) and time in ICU until discharge from ICU (bottom right).

Further, the LoS until a final outcome that will be experienced by only a proportion of patients is estimated with the NP-MCM and KM using both the complete and the reduced data sample. This is the case when estimating 5 key LoS: from admission into the hospital ward (HW) to admission into the ICU, from admission into HW to discharge (alive), from admission into HW to death, from admission into the ICU to discharge, and from admission into the ICU to death. In this case, KM (with both the complete and reduced samples) overestimates the time-to-event showing longer LoS than the NP-MCM. Interestingly, we found small differences between the NP-MCM estimates and the E estimates. The E estimator only takes into account data from patients who experienced the event; thus it cannot handle right censoring. Nevertheless, we find that the E estimator underestimates the time-to-event due to right censoring, showing shorter values of LoS. **Figure 1** shows NP-MCM, KM, and E estimators for the 5 key LoS analyzed. Details can be found in the Appendix, including plots of all the density function estimates for alternative visualization of the LoS in **Figure 6**.

Importantly, for the individual's probability of the medical event (admission from HW to ICU, and death or discharge from HW or ICU) we were able to show that not correcting for right censoring (i.e., using only individuals with the observed outcome) underestimates the true probability, as the event of the right-censored individuals could be recorded later in time. The NP-MCM can adjust to right censoring, providing more accurate estimates. This can be seen when comparing individual probabilities using NP-MCM and E estimators, as presented in **Table 1**.

**Table 1.** Estimated probabilities of the different medical events for the COVID-19 patients in Galicia (Spain) using NP-MCM and empirical estimators.

We then performed survival analysis using the NP-MCM estimator to assess if age and sex could play a role in the estimates of the time of hospitalization (both hospital ward and ICU) and the time in ICU. **Figure 2** shows that the duration times differ significantly between male and female patients, and between middle-aged (40y) and older (70y) patients. Particularly, we found that middle-aged female patients showed shorter LoS in both the institution of hospital and the ICU, while older females showed longer LoS in the ICU (but not in the hospital) compared to their male counterparts.

**Figure 2.** Generalized product-limit estimator [13] of the conditional survival function S(t|x) for the time of hospitalization, both in hospital ward and ICU, (top) and the time in ICU (bottom), incorporating the effect of the sex (male = black line, female = red line) and the ages 40y (left) and 70y (right) for all the COVID-19 hospitalized cases (n = 2453) in Galicia (Spain).

Finally, we implemented a COVID-19 outbreak simulation using the NP-MCM estimates for the COVID-19 patients in Galicia (Spain) and accounting for age and sex heterogeneity in the LoS. **Figure 3** shows the difference between considering age and sex in the estimated LoS or not. We found no large differences in the expected number of patients admitted to the hospital ward, regardless of age or sex. However, the unconditional distribution tends to overestimate the mean number of patients in the ICU with shorter stays, and to underestimate those with longer stays. The expected number of deaths was slightly higher for longer stays in the unconditional model than when age and sex were taken into account. Considering the age and sex of the patients did not yield a different number of discharges than when using the unconditional model. In general, the conditional model estimates longer stays (around 200 days in some cases) than the unconditional one. Furthermore, we compared the differences between the LoS in the HW and in the ICU for both the conditional and the unconditional models. **Figure 4** shows the time interval in which the expected number of patients is above the maximum bed capacity for a range of values between 15 and 75 beds in the HW and between 5 and 11 beds in the ICU. In summary, while no large differences are observed in the predicted HW beds demand, the conditional model that considers the age and sex of the patients gives much more accurate estimates of the ICU beds demand, reducing the number of days the number of ICU beds needed exceeds the capacity.

**Figure 3.** Simulated mean number of patients in hospital ward (top left), ICU (top right), deaths (bottom left) and discharges (bottom right), with distributions conditionally estimated depending on age and sex (blue) and unconditionally estimated, ignoring age and sex dependence (red)

**Figure 4.** Simulated number of days demand is above the capacity (threshold) of hospital ward (left) and ICU (right), when distributions are estimated conditionally depending on age and sex (black) and unconditionally estimated, ignoring age and sex dependence (red).

## Discussion

We applied a NP-MCM to estimate time-to-event and event probability using survival functions of key variables of hospital services, including length-of-stay in ICU and time to death or discharge. The proposed model outperformed the KM and the empirical estimators for computing the time to a final outcome that is not experienced by all patients. Importantly, the model can be adjusted for the use of covariates, which is significant when conditioning for known heterogeneity in estimating LoS. Particularly, our analysis demonstrates that adjusting for age and sex is crucial in accurately understanding ICU LoS and, in turn, forecasting bed demand.

Often studies with incomplete follow-up data on patients (called right censored data) choose to exclude these patients from the study altogether, which yields biased estimates [18]. Moreover, when forecasting hospital demand in (near) real time, information related to the most recent cases is not available, which again leads to right censored data. For instance, we showed that the empirical estimator introduced significant bias toward longer LoS for time from HW admission to ICU admission because it ignores patients discharged from the HW without ICU admission. Alternately, using information of patients without ICU admission by the end of the study period but ignoring that a proportion of patients will not require ICU yields biased estimates towards longer stays. The reason is that the KM estimator assumes that if the follow-up time was long enough, much longer stays would be observed (see **Figure 1**).

Our findings resonate with previous work: a recent systematic review has shown that median overall hospital stays ranged from 4 to 21 days outside of China [19], while our model estimated a median overall hospital stay of 11 days (IQR 7 – 19); the LoS for patients who died in the HW was generally shorter than those discharged alive (median of 7 days and 10 days respectively). In contrast, our estimates show a different trend with regards to ICU LoS, with similar median estimates for both death and discharged (15 days vs 14 days), again consistent with that reviewed by Rees et al [19]. Of note, to our knowledge only two studies have adjusted LoS by age, all showing increased LoS for increased age, which is consistent with our findings [20, 4]. Furthermore, as far as we know this is the first study showing the influence of sex in the LoS, which has important implications for predicting hospital demand (**Figure 2**). With regards to prediction models, some approach adjust estimates based on age [21, 22], while sex has generally been overlooked in hospital demand forecasting [23, 21, 7].

Noteworthy, multi-state models [24, 25] are a possible alternative method to our approach. Yet, formulation is not straightforward: the literature under flexible nonparametric conditions [26] and in the presence of covariates [27] is limited and deals only with estimating the conditional transition probabilities of the event covariates. In addition, selection of the smoothing parameter remains an open problem [28]. As a consequence, multi-state models were not used in this paper, but remain as a potential alternative approach.

Finally, we would like to highlight key limitations of our model: the lack of a parametric function limits interpretability to a great extent and complicates handling several covariates simultaneously [29]. Regarding the application of MCM, there must be good evidence that some individuals in the population will never experience the event of interest [30], and, the follow-up time must be long enough. Finally, data on patient comorbidities, which likely represents an important source of heterogeneity in the LoS, were not available for the analysis. Thus, more accurate estimates of the different LoS can be obtained if more complete datasets are available.

In summary, we implemented a NP-MCM that improved the standard survival methodology when estimating the LoS in the HW and in the ICU until final outcomes that will not happen for a proportion of patients. We also found that the LoS in the ICU is sensitive to age and sex, which in turn is relevant when forecasting hospital demand in real-time for public health response. We believe our proposed approach can be easily implemented in other settings and can provide more accurate estimates of COVID-19 health demand compared to previous methods.

**Acknowledgments:** We thank Marc Lipsitch, Iñaki López de Ullibarri and Rene Niehus for their valuable technical advice. We also thank the Dirección Xeral de Saúde Pública, Xunta de Galicia for providing the database.

**Funding**: ALC was sponsored by the BEATRIZ GALINDO JUNIOR Spanish from MICINN (Ministerio de Ciencia, Innovación y Universidades) with reference BGP18/00154. ALC, MAJ and RC acknowledge partial support by the MINECO (Ministerio de Economía y Competitividad) Grant MTM2014-52876-R (EU ERDF support included) and the MICINN Grant MTM2017-82724-R (EU ERDF support included) and partial support of Xunta de Galicia (Centro Singular de Investigación de Galicia accreditation ED431G/01 2016-2019 and Grupos de Referencia Competitiva CN2012/130 and ED431C2016-015) and the European Union (European Regional Development Fu–d - ERDF). PMD is a current recipient of the Grant of Excellence for postdoctoral studies by the Ramón Areces Foundation.

## References

1. World Health Organization (WHO). Coronavirus disease (COVID-19) Situation Report – 162.https://www.who.int/docs/default-source/coronaviruse/20200630-covid-19-sitrep-162.pdf?sfvrsn=e00a5466_2

2. Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. Projecting the transmission dynamics of SARS-CoV2-through the postpandemic period. *Science* 2020; **368**(6493): 860-868. DOI: 10.1126/science.abb5793

3. Grasselli G, Pesenti A, Cecconi M. Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy: Early Experience and Forecast During an Emergency Response. *JAMA* 2020; **323**(16):1545–1546. doi:10.1001/jama.2020.4031

4. Lewnard JA *et al.* Incidence, clinical outcomes, and transmission dynamics of severe coronavirus disease 2019 in California and Washington: prospective cohort study. *BMJ* 2020; **369**:m2205. https://doi.org/10.1136/bmj.m1923.

5. Thai P *et al.* Factors associated with the duration of hospitalization among COVID-19 patients in Vietnam: A survival analysis. *Epidemiol. Infect.* 2020; ***148***, E114. doi: 10.1017/S0950268820001259

6. Wang Z *et al*. Survival analysis of hospital length of stay of novel coronavirus (COVID-19) pneumonia patients in Sichuan, China. 2020b *medRxiv*, 2020-040720057299 (2020b). doi:10.1101/2020.04.07.20057299.

7. Grasselli G *et al*. Risk Factors Associated With Mortality Among Patients With COVID-19 in Intensive Care Units in Lombardy, Italy. *JAMA Intern Med*, 2020. doi:10.1001/jamainternmed.2020.3539

8. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assn.* 1958; 53(282):457–481.doi:10.2307/2281868

9. Boag JW. Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy. *J. R. Stat. Soc., Ser. B Stat. Methodol.*1949; **11**(1):15-53. http://www.jstor.org/stable/2983694

10. Dirección Xeral de Saúde Pública (General Directorate of Public Health), Xunta de Galicia (Autonomous Government of Galicia, NW Spain). https://www.sergas.es/Saude-publica

11. Maller RA, Zhou S. Survival Analysis with Long-Term Survivors. Chichester, U.K.: Wiley, 1996. doi: 10.1002/cbm.318

12. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

13. Beran R. Nonparametric regression with randomly censored survival data. Technical Report University of California, Berkeley; 1981.

14. Xu J, Peng Y. Nonparametric cure rate estimation with covariates. *Canad J Statist* 2014; **42**(1):1-17. https://doi.org/10.1002/cjs.11197

15. López-Cheda A, Cao R, Jácome MA, Van Keilegom I. Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Comput Stat Data An.* 2017a; **105**:144-165. doi:10.1016/j.csda.2016.08.002

16. López-Cheda A, Jácome MA, Cao R. Nonparametric latency estimation for mixture cure models. *TEST* 2017b; **26**(2):353-376. doi:10.1007/s11749-016-0515-1

17. López-de-Ullibarri I, López-Cheda A, Jácome MA. npcure: Nonparametric Estimation in Mixture Cure Models. 2019 R package version 0.1-4. https://CRAN.R-project.org/package=npcure

18. Lapidus N *et al*. Biased and unbiased estimation of the average lengths of stay in intensive care units in the COVID-19 pandemic. *medRxiv*, 2020; 2020–042120073916. doi:10.1101/2020.04.21.20073916

19. Rees EM *et al*. COVID-19 length of hospital stay: a systematic review and data synthesis. *medRxiv* 2020; https://doi.org/10.1101/2020.04.30.20084780

20. Wang L *et al*. Coronavirus disease 2019 in elderly patients: Characteristics and prognostic factors based on 4-week follow-up. *J. Infect*. 2020a; doi:10.1016/j.jinf.2020.03.019.

21. Moghadas SM et al. Projecting hospital utilization during the COVID-19 outbreaks in the United States. Proc Natl Acad Sci USA 2020; 117(16):9122-9126. doi:10.1073/pnas.2004064117

22. Li R *et al*. Estimated Demand for US Hospital Inpatient and Intensive Care Unit Beds for Patients With COVID-19 Based on Comparisons With Wuhan and Guangzhou, China. *JAMA Netw Open.*2020; **3**(5):e208297. doi:10.1001/jamanetworkopen.2020.8297

23. Wood RM, McWilliams CJ, Thomas MJ, Bourdeaux CP, Vasilakis C. COVID-19 scenario modelling for the mitigation of capacity-dependent deaths in intensive care. *Health Care Manag Sci*. 2020; **23**(3):315-324. doi:10.1007/s10729-020-09511-7

24. Andersen PK, Keiding N. Multi-state models for event history analysis. *Stat Methods Med Res* 2020; **11**(2):91-115. doi: 10.1191/0962280202SM276ra

25. Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C. Multi-state models for the analysis of time-to-event data. *Stat Methods Med Res* 2009; **18**(2):195-222. doi: 10.1177/0962280208092301

26. Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C. Nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Anal* 2006; **12**(3):325-344. doi:10.1007/s10985-006-9009-x

27. Meira-Machado L, de Uña-Álvarez J, Datta S. Nonparametric estimation of conditional transition probabilities in a non-Markov illness-death model. *Computation* Stat 2015; **30**(2):377-397. doi: 10.1007/s00180-014-0538-6

28. Zhang Z *et al*. Overview of model validation for survival regression model with competing risks using melanoma study data. *Ann Transl Med* 2018; **6**(16):325. doi: 10.21037/atm.2018.07.38

29. Bellman RE. Adaptive Control Processes. Princeton University Press, Princeton, NJ; 1961.

30. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1982; **38:**1041–46.

31. Li Q *et al*. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* (2020) ***382***(13):1199–1207 https://doi.org/10.1056/NEJMoa2001316

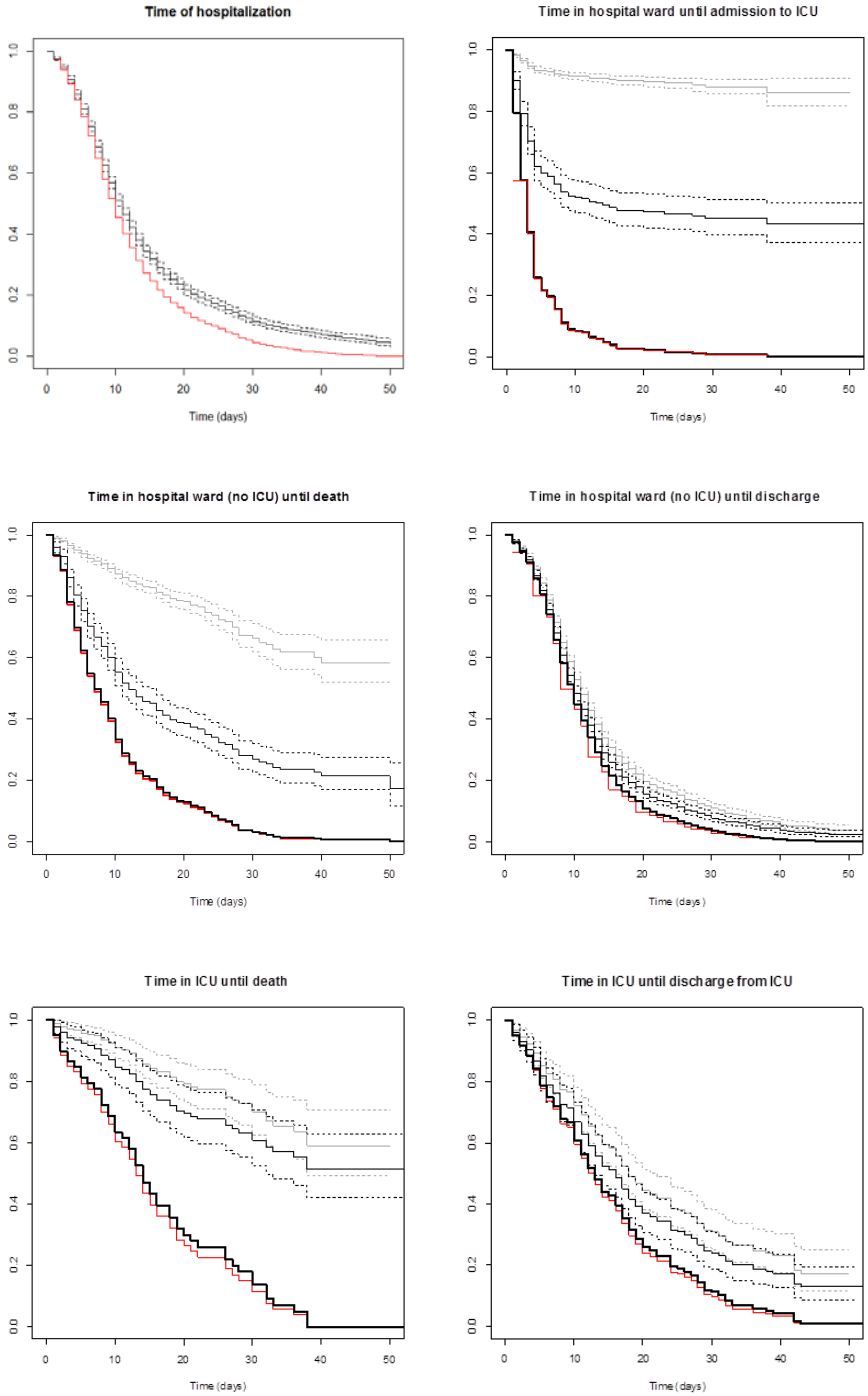


**Figure 1.** Estimates of the survival function of LoS using NP-MCM (thick black line), KM with the complete dataset (thin grey line), KM with the reduced dataset (thin black line) and the empirical E estimator (red line) for all the COVID-19 hospitalized cases (n = 2453) in Galicia (Spain), when the LoS is the time of hospitalization both in hospital ward and ICU (top left), time in hospital ward until admission to ICU (top right), time in hospital ward until death in hospital ward (middle left), time in hospital ward until discharge (middle right), time in ICU until death in ICU (bottom left) and time in ICU until discharge from ICU (bottom right).

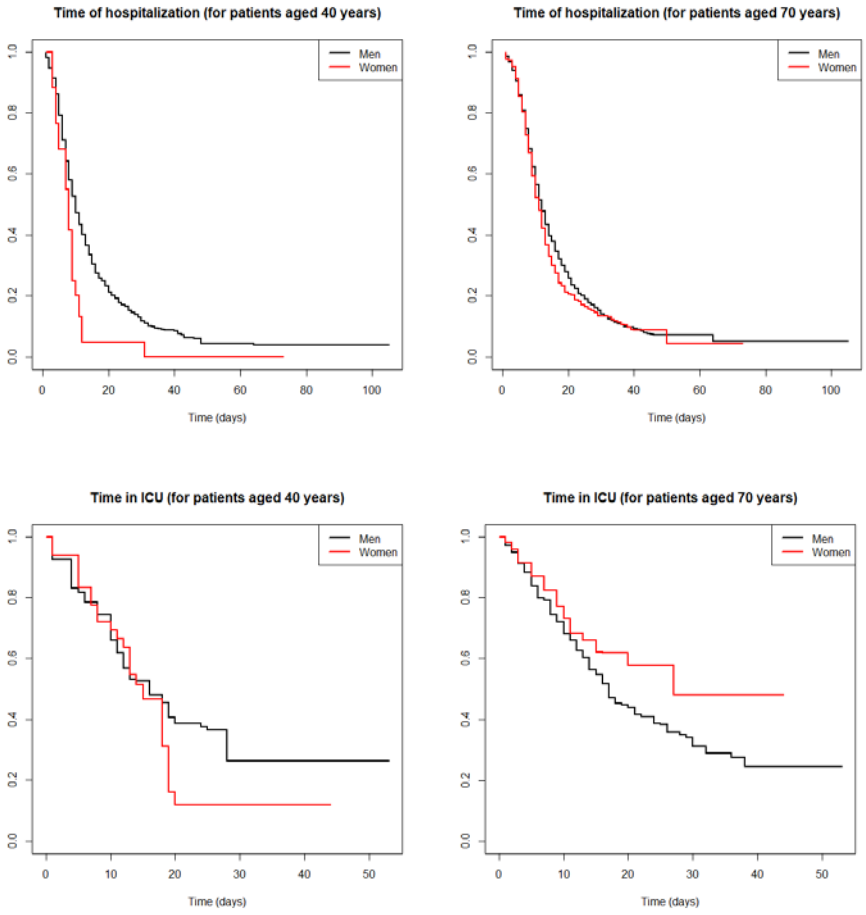


**Figure 2.** Generalized product-limit estimator [13] of the conditional survival function S(t|x) for the time of hospitalization, both in hospital ward and ICU, (top) and the time in ICU (bottom), incorporating the effect of the sex (male = black line, female = red line) and the ages 40y (left) and 70y (right) for all the COVID-19 hospitalized cases (n = 2453) in Galicia (Spain).

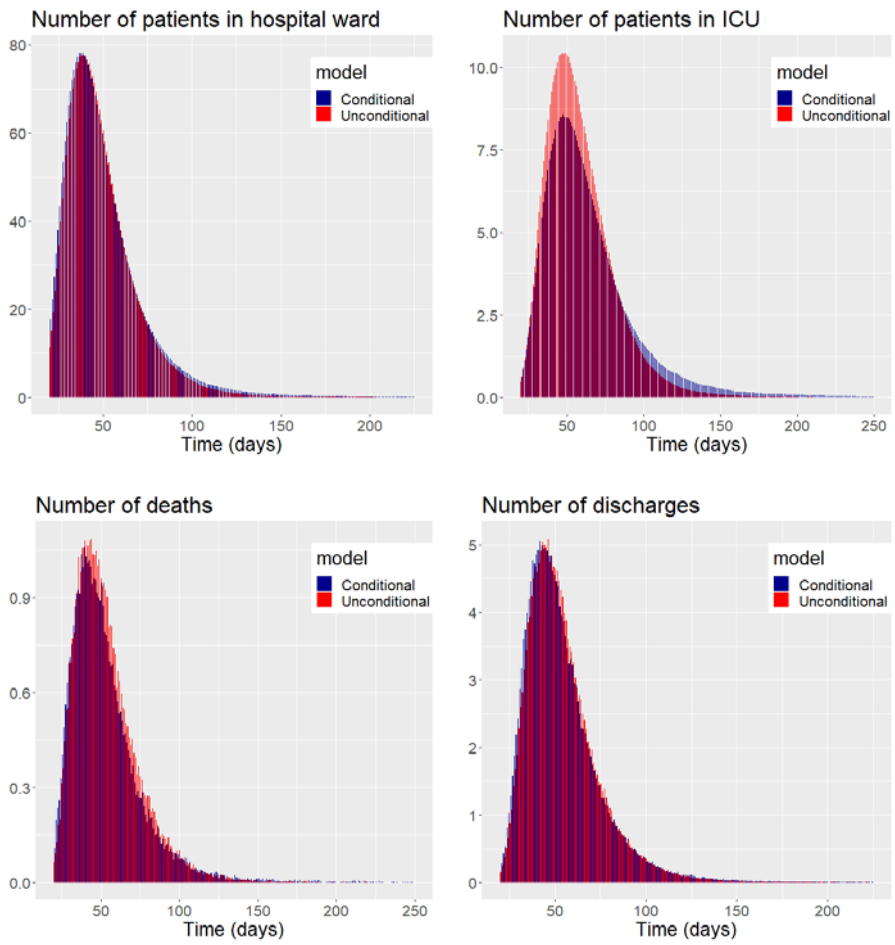


**Figure 3.** Simulated mean number of patients in hospital ward (top left), ICU (top right), deaths (bottom left) and discharges (bottom right), with distributions conditionally estimated depending on age and sex (blue) and unconditionally estimated, ignoring age and sex dependence (red).

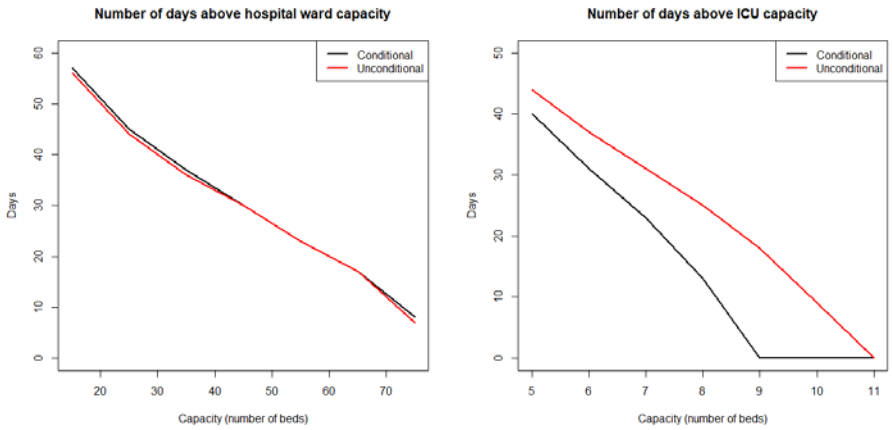


**Figure 4.** Simulated number of days demand is above the capacity (threshold) of hospital ward (left) and ICU (right), when distributions are estimated conditionally depending on age and sex (black) and unconditionally estimated, ignoring age and sex dependence (red).

**Table 1.** Estimated probabilities of the different medical events for the COVID-19 patients in Galicia (Spain) using NP-MCM and empirical estimators.

| | NP-MCM | Empirical |
|---|---|---|
| Need for ICU | 0.0845 | 0.0828 |
| Death in HW | 0.1561 | 0.1503 |
| Discharge from HW | 0.7953 | 0.7503 |
| Death in ICU | 0.2222 | 0.1963 |
| Discharge from ICU | 0.6820 | 0.6481 |

HW: Hospital ward; ICU: Intensive care unit

# Appendix for

## Estimating COVID-19 hospital demand using a non-parametric model: a case study in Galicia (Spain)

**Authors**: Ana López-Cheda[1], María-Amalia Jácome*[1], Ricardo Cao[2], Pablo M De Salazar[3]

**Affiliations**
1. Universidade da Coruña, CITIC, MODES, A Coruña, Spain
2. Universidade da Coruña, CITIC, ITMATI, MODES, A Coruña, Spain
3. Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard School of Public Health, Boston, US

* Corresponding author: María-Amalia Jácome.
Faculty of Science, Rúa da Fraga 10, 15008 A Coruña (Spain)
**Email:** maria.amalia.jacome@udc.es.

**ORCIDs**: Ana López-Cheda (https://orcid.org/0000-0002-3618-3246)[1], María-Amalia Jácome (https://orcid.org/0000-0001-7000-9623)[1], Ricardo Cao (https://orcid.org/0000-0001-8304-687X)[2], Pablo M De Salazar (https://orcid.org/0000-0002-8096-2001)[3]

**This Appendix includes:**

Appendix text
Figures 5 to 9
Tables 2 to 4
References
Script (R) for reproducing the results

**Details on the dataset**

From a total of 10454 reported cases, 2484 were admitted to the hospital, though 31 of them were discharged on the same day. Among the 2453 patients admitted to the hospital for at least one day, 281 needed care in the ICU (11.45%), and 270 stayed in the ICU for at least one day. These 270 patients with long stays in the ICU can be divided into 197 patients admitted from the hospital ward, and 73 admitted directly from the emergency service. On May 7, 2020, 57 of the long-stay ICU patients had died, 119 had been discharged to the hospital ward, and 43 were still in the ICU. **Figure 5** includes a flowchart related to the database. For the distribution of cases for different ages and sex, see **Table 2**.

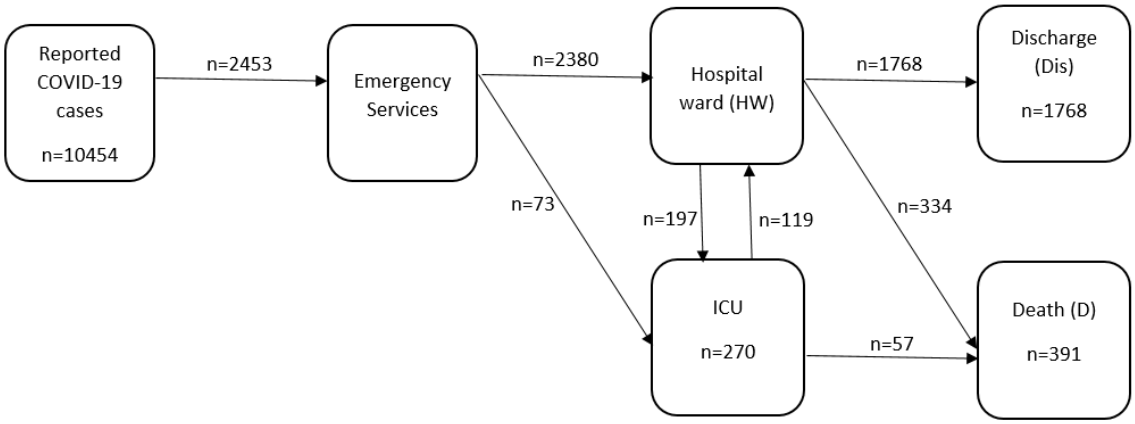


**Figure 5.** Flowchart of the confirmed COVID-19 cases reported in Galicia (Spain) from March 6 to May 7, 2020.

**Table 2.** Distribution of the total number of reported COVID-19 cases, and the number of reported cases hospitalized in Galicia (Spain) from March 6 to May 7, 2020.

| | Reported | | Hospitalized | |
|---|---|---|---|---|
| Age | Women | Men | Women | Men |
| +90 | 397 | 146 | 127 | 71 |
| 80-89 | 743 | 495 | 307 | 286 |
| 70-79 | 736 | 758 | 258 | 401 |
| 60-69 | 919 | 735 | 190 | 274 |
| 50-59 | 1150 | 705 | 118 | 161 |
| 40-49 | 1096 | 609 | 96 | 84 |
| 30-39 | 724 | 378 | 45 | 26 |
| 20-29 | 402 | 205 | 17 | 9 |
| 10-19 | 81 | 83 | 4 | 4 |
| 0-9 | 34 | 58 | 2 | 2 |
| Total | 6282 | 4172 | 1164 | 1320 |

**Details on NP-MCM, KM and E estimates of time-to-event when there is a group of patients who will not experience the final outcome: ICU admission from the hospital ward, death or discharge**

The estimator of S(t) in (eq2) is used to estimate the probability, p, of the event and the distribution of the times-to-event $S_0(t)$ using the relationships in (eq1) for the following lengths of stay.

*Time in hospital ward (HW) until admission to ICU*

The goal is to estimate the probability that a patient in HW will need admission to ICU, and the distribution of the LoS in HW of those patients. The observations are $\{(t_i, d_i, x_i), i = 1,\ldots,n\}$ with $t_i$ the observed LoS in HW of all the patients, $d_i$ indicates if the patient was admitted to ICU, and $x_i$ the indicator of whether the admission to ICU was not observed because it will never happen because the patient died in HW or was discharged.

There were 2453 COVID-19 patients admitted to the hospital. In order to study the time in HW until ICU, we worked with the n = 2380 patients who were admitted first to HW, discarding the 73 patients who went to ICU directly from the emergency service. In the group of n = 2380 patients in HW, 197 of them required ICU. This gives an estimated empirical (E) probability of need for ICU $p_{emp}$ = 197/2380 = 0.0828. But note that some of the patients still in HW at the end of the study would be admitted to ICU eventually, so the real probability is expected to be larger. NP-MCM approach estimates that probability to be $p_{NP\text{-}MCM}$ = 0.0845. The classical KM estimator considers n = 2380 patients in HW where 197 patients with admission to ICU is observed. This classical KM assumes that all the patients who had been admitted to HW will experience the event (admission to ICU) if followed for long enough, overestimating the LoS. This bias is partially corrected by the improved KM estimator, which takes into account that 1638 patients were discharged without ICU, and 328 died before being admitted to ICU. So it considers only n = 2380 – 1638 – 328 = 414 patients in HW with 197 patients where the event (admission to ICU) is observed. It still biases towards larger LoS, as patients still in HW by the end of the study are assumed to require ICU sometime in the future. The empirical estimator considers only n = 197 patients who were admitted to ICU, disregarding the information from the other right censored patients.

*Time in hospital ward (HW) until death in HW*

The aim is to estimate the probability that a patient will die in HW, and the distribution of the LoS in HW of those patients. The observations are $\{(t_i, d_i, x_i), i = 1,\ldots,n\}$ with $t_i$ the observed LoS in HW of all the patients, $d_i$ indicates if the patient died in HW, and $x_i$ the indicator of whether the patient will not die in HW since he/she was discharged alive.

There were 2453 COVID-19 patients admitted to the hospital (into a hospital ward or the ICU). To study the time in HW until death, we worked with the n = 2183 patients who never required admission to ICU. In that group, 328 patients died, which gives an estimated empirical probability of death $p_{emp}$ = 328/2183 = 0.1503. However some of the 2183 patients were still in HW at the end of the study, and they might die eventually, so the probability of death in HW is expected to be larger. NP-MCM approach estimates that probability to be $p_{NP\text{-}MCM}$ = 0.1561.

Note that 1638 patients will never die in HW because they have been discharged; they are the known “cures” from death in HW. The classical KM estimator considers n = 2183 patients in HW with 328 observed events. The improved KM estimator takes into account that 1638 patients were discharged alive. So it considers only n = 2183 – 1638 = 545 patients in HW with 328 patients where the event (death) is observed. The empirical estimator considers only the n = 328 patients whose event is observed, that is, those who died in HW, disregarding the information from the other patients.

*Time in hospital ward (HW) until discharged without ICU*

The goal is to estimate the probability that a patient in HW will be discharged without requiring ICU, and the distribution of the LoS in HW of those patients. The observations are $\{(t_i, d_i, x_i), i = 1,\ldots,n\}$ with $t_i$ the observed LoS in hospital ward of all the patients, $d_i$ indicates if the patient was discharged without need for ICU, and $x_i$ the indicator of whether discharge will not be observed because the patient died before that event happened.

To study the time in HW until discharge, we worked with the n = 2183 patients in HW who did not need intensive care. In that group, 1638 were discharged, so the empirical estimator of the probability of discharge from HW without need for ICU is $p_{emp}$ = 1638/2183 = 0.7503. However there were patients still in HW at the end of the study, and many of them are expected to be discharged without admission to ICU, so the true probability of discharge from HW without ICU should be larger than $p_{emp}$ =0.7503. The NP-MCM estimator of that probability is $p_{NP\text{-}MCM}$ = 0.7953.

Note that 328 of the 2183 patients in HW will never be discharged because they died. They are the known "cures" from discharge. The classical KM estimator considers the n = 2183 patients in HW with 1638 patients where the event (discharge) is observed. The improved KM estimator takes into account that 328 patients died and will never be discharged. So it considers only n = 2183 – 328 = 1855 patients in HW with 1638 patients discharged. The empirical estimator considers only the n = 1638 patients discharged from HW, disregarding the information from the other patients.

*Time in ICU until death in ICU*

The objective is to estimate the probability for a patient in ICU of dying, and the distribution of the LoS in ICU of those patients. The observations are $\{(t_i, d_i, x_i), i = 1,\ldots,n\}$ with $t_i$ the observed time in ICU of all the patients, $d_i$ indicates if the patient died in ICU, and $x_i$ the indicator of whether the patient was discharged alive from ICU.

There were n = 270 patients admitted to ICU, and 53 of them died in ICU, so the empirical probability of death in ICU is $p_{emp}$ = 53/270 = 0.1963. But in this group of n = 270 patients there were 42 patients still in ICU at the end of the study, and 52 in HW discharged from ICU who might need ICU again. Note that any patient within these two groups might die in ICU eventually, so the number of deaths in ICU for these 270 patients is expected to be larger than the 53 observed deaths. As a consequence, the true probability of death in ICU should be larger than $p_{emp}$ = 53/270 = 0.1963. The NP-MCM estimation of this probability is $p_{NP\text{-}MCM}$ = 0.2222.

To study the time in ICU until death, the NP-MCM estimator takes into account that some of the n = 270 patients in ICU will never die in ICU because they have been discharged from hospital (119) or died in HW after leaving ICU (4); they are the known "cures" from death in ICU. The classical KM estimator considers the n = 270 patients in ICU with n = 53 patients where the event (death) is observed. The improved KM estimator considers only n = 270 – 4 - 119 = 147 patients in ICU with 53 observed deaths. Finally, the empirical estimator considers only the n = 53 patients who died in ICU, disregarding the information from the other right censored times.

*Time in ICU until discharged from ICU to HW*

The goal is to estimate the probability that a patient in ICU will be sent back to the hospital ward, and the distribution of the times in ICU of those patients. The observations are $\{(t_i, d_i, x_i), i = 1,\ldots,n\}$ with $t_i$ the time in ICU of all the patients, $d_i$ indicates if the patient was discharged from ICU, and $x_i$ the indicator of whether the patient died in ICU.

There were n = 270 patients who required ICU. To estimate the probability of discharge from ICU, the empirical estimator considers the 175 patients who were discharged from ICU (4 dead in HW after ICU, 52 still in HW and 119 discharged at home at the end of the study). This yields an estimated probability of discharge from ICU of $p_{emp}$ = 175/270 = 0.6481. But some of the 42 patients still in ICU at the end of the study (53 patients died in ICU) may be discharged, so the real probability of discharge from ICU is expected to be slightly larger than 0.6481. NP-MCM approach estimates that probability to be $p_{NP\text{-}MCM}$ = 0.6820.

A total of 53 patients in ICU will never be discharged from ICU because they died in ICU; they are the known "cures" from discharge. The classical KM estimator considers the n = 270 patients in ICU with 175 observed events (discharge from ICU). The improved KM estimator takes into account that 53 patients died in ICU so they will never be discharged from ICU. It considers only n = 270 - 53 = 217 patients in ICU with 175 observed discharges. Finally, the empirical estimator considers only the n = 175 patients who have been discharged from ICU, disregarding the information from the other right censored patients.

**Table 3** shows the NP-MCM estimations of the time of hospitalization (which considers time in HW plus time in ICU) and the time in ICU for male and female patients, considering ages 40 and 70 years.

**Table 3.** NP-MCM estimation of the mean, median and IQR of the duration of hospitalization (which considers time in HW plus time in ICU) and time in ICU, computed with the COVID-19 cases hospitalized in Galicia (Spain) from March 6 to May 7, 2020.

| | **Duration of hospitalization** | | | **Time in ICU** | | |
|---|---|---|---|---|---|---|
| | **Mean[1]** | **Median** | **IQR** | **Mean[1]** | **Median** | **IQR** |
| Total | 16.80 | 11 | 7-19 | 23.94 | 17 | 9-38 |
| Women | 13.92 | 11 | 7-20 | 21.76 | 17 | 8-38 |
| Men | 18.41 | 10 | 6-17 | 23.88 | 18 | 10-NA |
| 40 years | 15.36 | 9 | 6-16 | 21.82 | 16 | 8-28 |
| 70 years | 19.71 | 12 | 8-20 | 30.14 | 30 | 9-NA |
| Women 40y | 8.47 | 8 | 5-10 | 16.56 | 15 | 8-19 |
| Men 40y | 16.84 | 10 | 6-18 | 23.58 | 16 | 8-NA |
| Women 70y | 16.63 | 11 | 7-17 | 28.12 | 27 | 10-NA |
| Men 70y | 19.60 | 12 | 7-21 | 24.28 | 17 | 8-38 |

[1] Underestimate

## Density estimations for the different LoS.

The density functions in **Figures 6 - 8** correspond to the survival estimates in **Figures 1, 2** and **9** respectively. Both the NP-MCM and empirical estimators yield proper survival functions (they go down to zero as time *t* increases), so the corresponding density functions are proper (area equal to 1). Note that, however, when there are individuals who will not experience the final outcome, the KM is wrongly specified and the curves reach a plateau at the largest observed time, $t^1_n$. This has two important implications in the density function corresponding to the KM curve: (a) the area is not 1 but lower (the difference between 1 and the plateau), so it should not be used for comparisons to proper density functions as those corresponding to the NP-MCM and E estimators; and (b) the KM curves estimate a large percentage of individuals (% = the value of the plateau) experiencing the event after the largest observed time, $t^1_n$, but the density function after $t^1_n$ is zero. This zero value of the density function should not be interpreted as no events after that time $t^1_n$ but simply lack of knowledge.

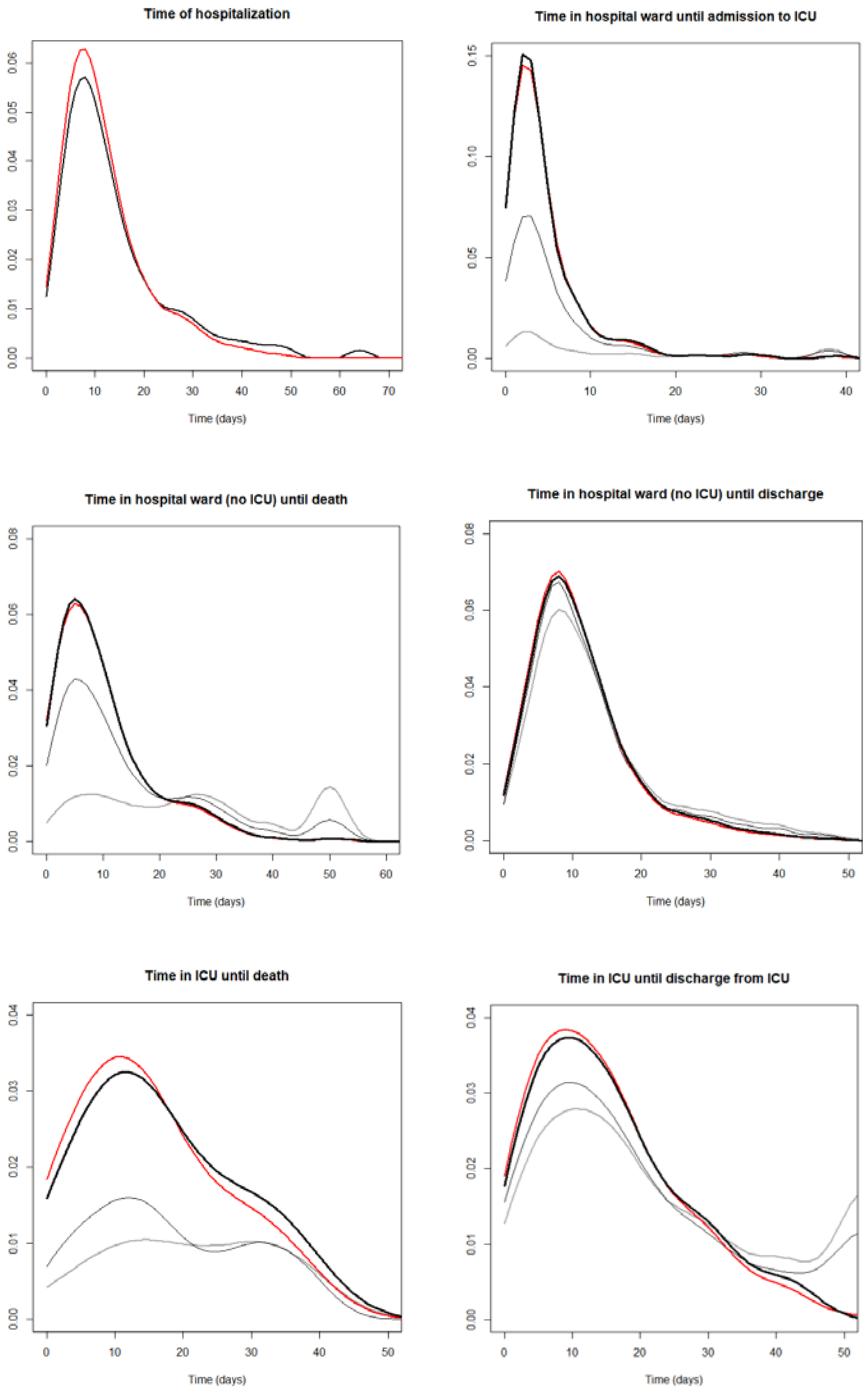


**Figure 6.** Estimates of the density function of LoS using NP-MCM (thick black line), KM with the complete dataset (thin grey line), KM with the reduced dataset (thin black line) and the empirical E estimator (red line) for all the COVID-19 hospitalized cases (n = 2453) in Galicia (Spain), when the LoS is the time of hospitalization (top left), time in hospital ward until admission to ICU (top right), time in hospital ward until death in hospital ward (middle left), time in hospital ward until discharge (middle right), time in ICU until death in ICU (bottom left) and time in ICU until discharge from ICU (bottom right).

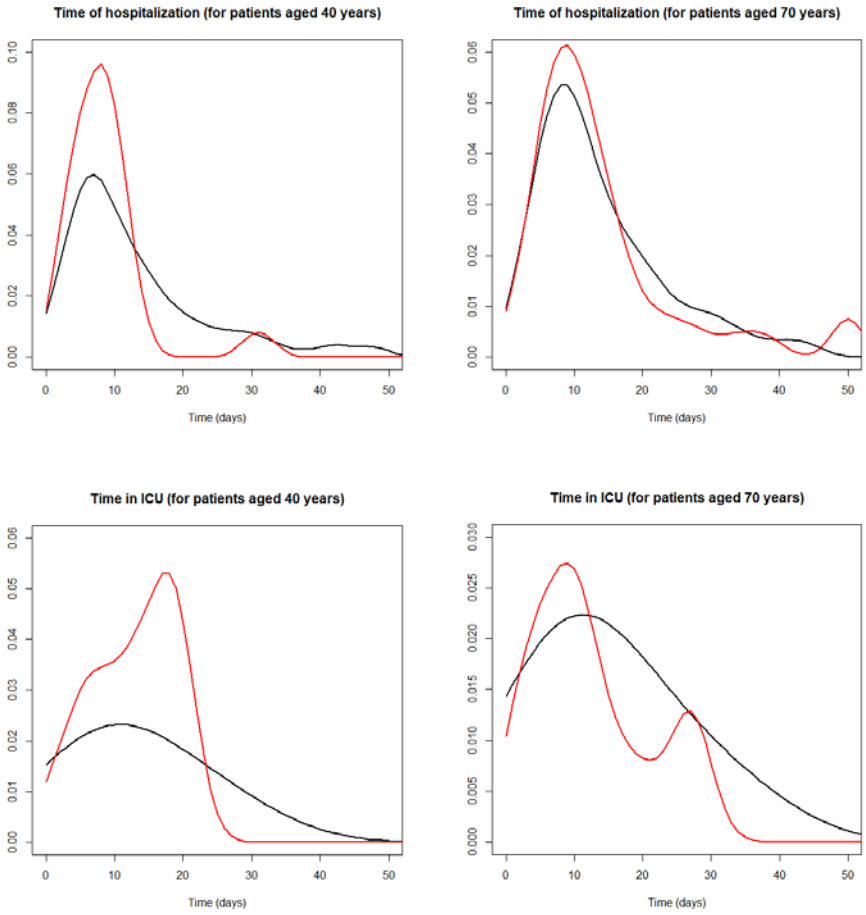


**Figure 7.** Estimates of the density function of the times of hospitalization which considers HW plus ICU (top) and time in ICU (bottom), incorporating the effect of the sex (male = black line, female = red line) and the ages 40y (left) and 70y (right) for all the COVID-19 hospitalized cases (n = 2453) in Galicia (Spain).

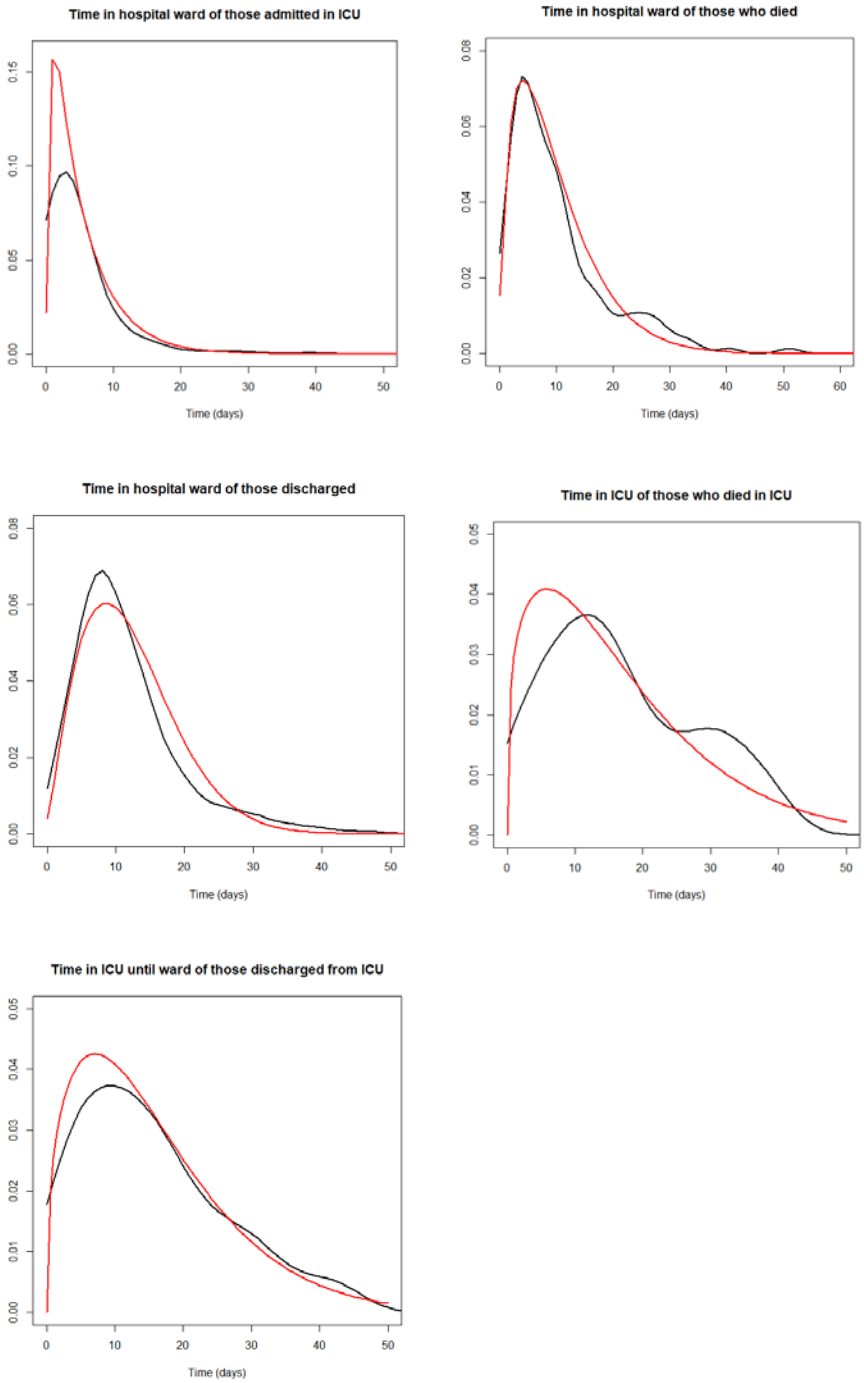


**Figure 8.** Estimation of the density function of different times-to-event using the NP-MCM (black) and Weibull distribution (red) unconditionally without taking into account sex and age of the patients, of the LoS in hospital ward until admission in ICU (top left), LoS in hospital ward until death (top right), LoS in hospital ward until discharge (middle left), LoS in ICU until death (middle right) and LoS in ICU until discharge to hospital ward (bottom left) for all the COVID-19 hospitalized cases (n = 2453) in Galicia (Spain).

**Model for simulating outbreak**

We considered a simulated outbreak with N = 1000 infected individuals. We assumed that the infection times, $I_i$, i = 1, …, N, followed a log-Normal distribution, *Log-N($\mu$, $\sigma$),* with $\mu$ = 3.3 days and $\sigma$ = 0.5 days [31]. For every i = 1, …, N, we simulated the sex $G_i$ (0 = male, 1 = female) and the age $A_i$ (years) of the *i-th* infected individual using the real distributions of the reported COVID-19 cases in Galicia on May 7, 2020 (for details in case counts see **Table 2**).

We defined H ∈ {1, …, N} as the set of indices corresponding to infected subjects admitted in hospital. The trajectory of every patient *i* ∈ H was obtained by simulating the transitions between states of the state space *S* = {HW, ICU, D, Dis}, where *D* (death) and *Dis* (discharge) are terminal states, using the NP-MCM estimates of the probabilities $p_{j,k}$ for j,k ∈ S. The duration times in states in *S* until transition to another state in *S* were also simulated using the Weibull distributions that best fitted the NP-MCM survival estimates.

The proposed model was used to perform a Monte Carlo simulation as follows. For every patient *i* = 1, …, N with age $A_i$ and sex $G_i$ the probability of admission in hospital $\pi(A_i, G_i)$ is estimated from the reported and hospitalized cases in **Table 2**. Based on a U(0,1) random variate, $U_i$, patient *i* is included in the set *H* of patients to be admitted into the hospital if $U_i \leq \pi(A_i, G_i)$. This gave us the set *H*. The time since infection until hospital admission $T_i$, of a patient *i* ∈ H was simulated from a normal distribution $N(\mu_i, \sigma_i)$ with $\mu_i = 12 - 0.05A_i$ days and $\sigma_i = 1$.

When a patient is admitted into the hospital, the probability of going directly to ICU is $p_{H,ICU}$ = 0.03, while the probability of staying in the hospital ward first is $p_{H,HW}$ = 0.97. In the simulated model conditioned on the age and sex of the patient, of those admitted in hospital ward, the probability of death without going to ICU is $p_{HW,D}(i) = 0.005\exp(0.045Ai)$ and the time (days) to death follows a Weibull age-dependent and sex-dependent distribution, $W(\alpha_{HW,D}(i), \lambda_{HW,D}(i))$, with parameters $\alpha_{HW,D}(i) = 1.4 - 0.2Gi$ and $1/\lambda_{HW,D}(i) = 20\exp(-0.008Ai)$ for every *i* ∈ H. The probability that a patient admitted in hospital ward finally has to enter ICU is $p_{HW,ICU}$ = 0.085, with the time (days) since hospital ward admission to ICU admission generated from a Weibull distribution $W(\alpha_{HW,ICU}(i), \lambda_{HW,ICU}(i))$ with $\alpha_{HW,ICU}(i) = 2.75 - 0.025Ai$ and $1/\lambda_{HW,ICU}(i) = 2.5\exp(0.02A_i)$. As a consequence, the probability that a patient who was admitted to the hospital ward becomes discharged without entering ICU is $p_{HW,Dis}(i) = 0.915 - 0.005\exp(0.045A_i)$. The time (days) since hospital ward admission to discharge follows a Weibull distribution $W(\alpha_{HW,Dis}(i), \lambda_{HW,Dis}(i))$, with $\alpha_{HW,Dis}(i) = 1.75\ (75 + 0.5A_i - 11G_i) / 100$ and $1/\lambda_{HW,Dis}(i) = 13\ (-2.5 + 1.5A_i - 7.5G_i) / 100$. After being admitted in ICU a patient may die, with probability $p_{ICU,D}(i) = 0.0067\exp((0.045 - 0.01G_i)A_i)$ or be transferred back to hospital ward, with probability $p_{ICU,HW}(i) = 1 - p_{ICU,D}(i)$. Time from admission into ICU to death follows a Weibull distribution $W(\alpha_{ICU,D}(i), \lambda_{ICU,D}(i))$, with parameters $\alpha_{ICU,D}(i) = 0.8\exp(0.009 + A_i)$ and $1/\lambda_{ICU,D}(i) = 30\exp(-0.012A_i)$ for every *i* ∈ H. The distribution of the time since admission into ICU until return to ward is again Weibull $W(\alpha_{ICU,HW}(i), \lambda_{ICU,HW}(i))$, with $\alpha_{ICU,HW}(i) = 1.6\ (1 + G_i)\exp(-0.003A_i(1+4G_i))$ and $1/\lambda_{ICU,HW}(i) = 20\exp(-0.003A_i(1 - G_i) - 0.22G_i)$ for every *i* ∈ H. A summary of the considered Weibull parameters is presented in **Table 3** (see also **Figures 8** and **9**). Note that for the different Weibull distributions, the shape parameters $\alpha(i)$ are truncated to be higher than 0.5, whereas the scale parameters $1/\lambda(i)$ are truncated to be higher than 1. All the estimated probabilities $p_{j,k}$ for j,k ∈ S are truncated to fall between 0.05 and 0.95.

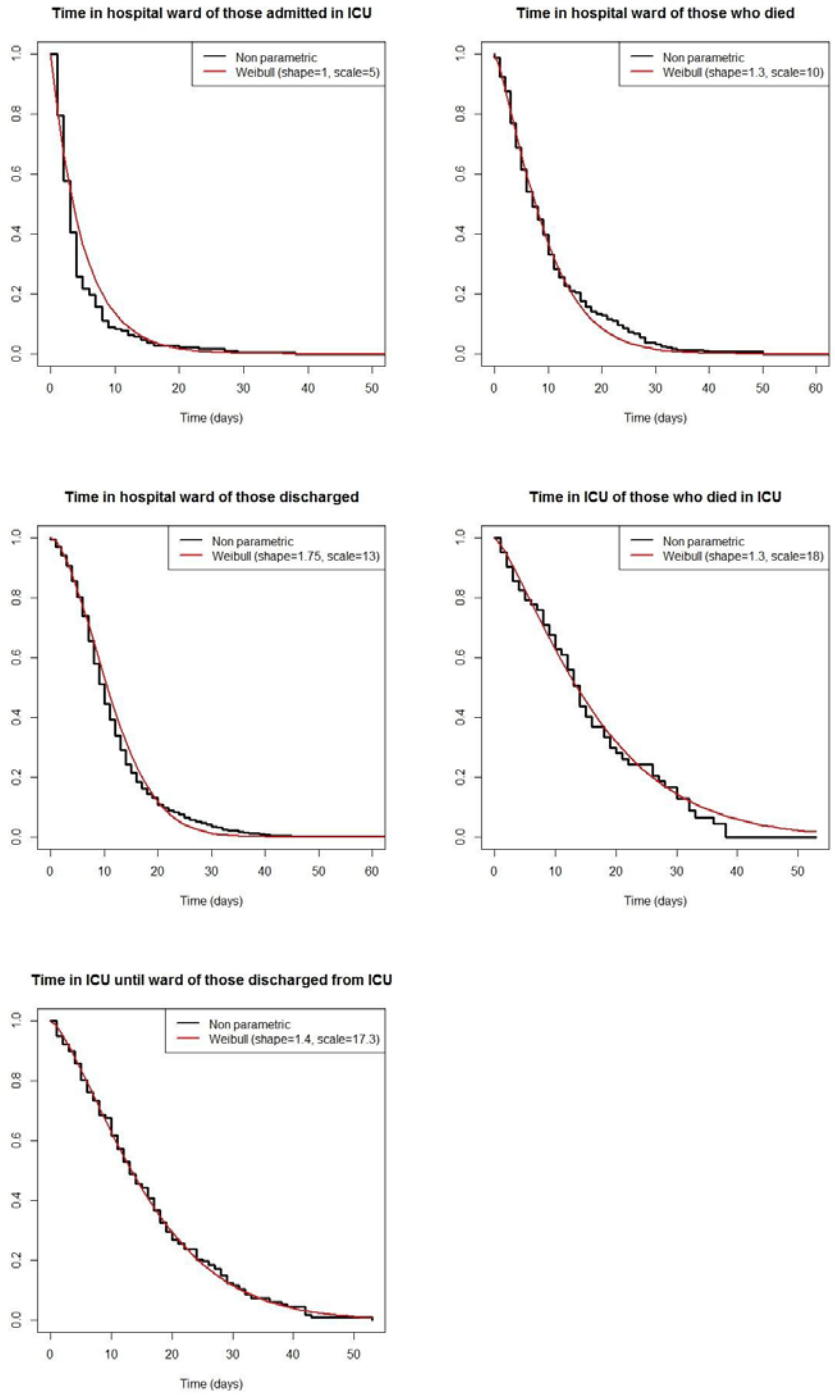


**Figure 9** Estimation of the survival function of different times-to-event using the NP-MCM (black) and Weibull distribution (red) unconditionally without taking into account the age and sex of the patients, of the LoS in hospital ward until admission into ICU (top left), LoS in hospital ward until death (top right), LoS in hospital ward until discharge (middle left), LoS in ICU until death (middle right) and LoS in ICU until discharge to hospital ward (bottom left) for all the COVID-19 hospitalized cases (n = 2453) in Galicia (Spain).

**Table 4**. Parameters of the Weibull distribution fitted to the different times-to-event, based on the hospitalized COVID-19 patients in Galicia (Spain) from March 6 to May 7, 2020.

| | **Model age (*Ai*) and sex (*Gi*) dependent** | | **Unconditional** | |
|---|---|---|---|---|
| **Times** | $\alpha(i)$ | $1/\lambda(i)$ | $\alpha$ | $\lambda$ |
| HW to ICU | $2.75 - 0.025A_i$ | $2.5\exp(0.02A_i)$ | 1 | 5 |
| HW to death | $1.4 - 0.2G_i$ | $20\exp(-0.008A_i)$ | 1.3 | 10 |
| HW to discharge | $1.75(75 + 0.5A_i - 11G_i)/100$ | $13(-2.5 + 1.5A_i - 7.5G_i)/100$ | 1.75 | 13 |
| ICU to death | $0.8\exp(0.009 + A_i)$ | $30\exp(-0.012A_i)$ | 1.3 | 18 |
| ICU to HW | $1.6(1 + Gi)\exp(-0.003A_i(1 + 4G_i))$ | $20\exp(-0.003A_i(1 - G_i) - 0.22G_i)$ | 1.4 | 17.3 |

HW: Hospital ward; ICU: Intensive care unit

The effect of ignoring the dependence on age and sex can be shown by simulating an alternative model where all the probabilities and time-to-event distributions do not depend on these variables. More specifically, the probabilities of death, discharge and admission to ICU in a hospital ward are $p_{HW,D} = 0.15$, $p_{HW,Disc} = 0.795$ and $p_{HW,ICU} = 0.085$ respectively. The probabilities of death in ICU and discharge from ICU are $p_{ICU,D} = 0.24$ and $p_{ICU,HW} = 0.76$. The shape and scale parameters of the Weibull distributions, which no longer depend on age nor sex, are specified in **Table 4**.

```
# This script contains the code for estimating:

# 1. The final outcome may happen for all the individuals
#    Example: Duration of hospitalization, time in ICU, etc
#    S(t)  : survival function (Kaplan and Meier, 1958)
#    S(t|x): survival function conditioned on x (Beran, 1981)

# 2. The outcome is not experienced for a subgroup of individuals,
#    some of them clearly identified as being event-free, cure
partially known
#    Example: Length of stay in hospital ward until admission in
ICU/discharged alive/death
#    p:     probability of experiencing the event (Safari et al, 2020)
#    S0(t): survival function of the individuals experiencing the event
(Safari et al, 2020)


#############################################################################
# 1. THE FINAL OUTCOME MAY HAPPEN FOR ALL THE INDIVIDUALS
#-------------------------------------------------------------
#    S(t)  : survival function (Kaplan and Meier, 1958)
#-------------------------------------------------------------

# Data frame - The observations are ordered based on the times Ti.
# time:   observed time to event
# status: indicator of whether the final outcome has been observed
data.real <- as.data.frame(cbind(time, status))

library(survival)
km_fit <- survfit(Surv(time, status) ~ 1, data = data.real)
km_fit$time   # Observed times
km_fit$surv   # Survival function S(t) evaluated at the observed times

#-------------------------------------------------------------
#    S(t|x): survival function conditioned on x (Beran, 1981)
#-------------------------------------------------------------
# Data frame - The observations are ordered based on the times Ti.
data.real <- as.data.frame(cbind(sex, age, time, status))
# sex:    sex of the individual
# age:    age of the individual
# time:   observed time to event
# status: indicator of whether the final outcome has been observed


# Covariate SEX: estimation of S(t|x) when x = 0 (men) and x = 1
(women).

library(survival)
km_fit.men <- survfit(Surv(time, status) ~ 1, data = data.real, subset
= sex == 0)
km_fit.men$time   # Observed times
km_fit.men$surv   # Survival function S(t) evaluated at the observed
                  # times

km_fit.women <- survfit(Surv(time, status) ~ 1, data = data.real,
subset = sex == 1)
km_fit.women$time   # Observed times
```

```
km_fit.women$surv   # Survival function S(t) evaluated at the
                    # times


# Covariate AGE: estimation of S(t|x) when x = 40 and x = 70.

library(npcure)
# Values of x = age where the survival function S(t|x) is estimated.
grid.age <- c(40, 70)
b_fit <- beran(x = age, t = time, d = status, dataset = data.real,
               x0 = grid.age,
               conflevel = 0.95,
               cvbootpars = controlpars(hbound = c(0.2, 2), hl = 100))

# Survival function for age = 40 years:
S.40 <- b_fit$S$x40
# 95% confidence band
S.40.lower <- b_fit$conf$x40$lower
S.40.upper <- b_fit$conf$x40$upper

# Survival function for age = 70 years:
S.70 <- b_fit$S$x70
# 95% confidence band
S.70.lower <- b_fit$conf$x70$lower
S.70.upper <- b_fit$conf$x70$upper


# Covariates AGE and SEX: estimation of S(t|x) when x = 40 and male

grid.age <- 40
data_men <- subset(data.real, sex == 1)
b_fit_men <- beran(x = age, t = time, d = status, dataset = data_men,
                   x0 = grid.age,
                   conflevel = 0.95,
                   cvbootpars = controlpars(hbound = c(0.2, 2), hl = 100))
# Survival function for age = 40 years and sex = male:
S.men.40 <- b_fit_men$S$x40
# 95% confidence band
S.men.40.lower <- b_fit_men$conf$x40$lower
S.men.40.upper <- b_fit_men$conf$x40$upper


#########################################################################
# 2. THE OUTCOME IS NOT EXPERIENCED FOR A GROUP OF INDIVIDUALS
# The survival function is S(t) = (1 - p) + p S0(t)
# p:     probability of experiencing the event
# S0(t): survival function of the individuals experiencing the event

# Data frame - The observations are ordered based on the times Ti:
# time:   observed time to event
# status: indicator of whether the final outcome has been observed
# cure:   indicator of whether the individual is known not to
experience the event (cured)
data.real <- as.data.frame(cbind(time, status, cure))
```

```
#=========================================================================
# Function that computes the nonparametric (NP) survival estimator
# when cure is partially known (CPK)
#=========================================================================
S_NPCPK <- function (data=data) {

  N <- nrow(data)

  # The observations are ordered based on the times Ti
  data.ot <- data[order(data[, 1]), ]
  t  <- data.ot[,1]
  d  <- data.ot[,2]
  nu <- data.ot[,3]
  cum.nu <- cumsum(nu)  # Number of known cures up to time Ti

  S <- rep(1, N)

  for (i in 2:N) {
    if(d[i]==0) {S[i] <- S[i-1]}
    if(d[i]==1) {S[i] <- S[i-1] * (1 - 1/(N - i + 1 + cum.nu[i-1]))}}

  p <- 1 - min(S)

  return(list(S, p, t))
}

#=========================================================================
# Survival estimation
# S[[1]]: survival function S(t) evaluated at the observed times
# S[[2]]: probability (1 - p) of not experiencing the event
# S[[3]]: observed times Ti

S <- S_NPCPK(data.real)

# p : probability of experiencing the final outcome
p <- 1 - S[[2]]

# S0(t): Survival function of the individuals experiencing the event
S0 <- (S[[1]] - (1 - p))/p
#=========================================================================
```