Competing risk modelling for in-hospital length of stay

Juan Carlos Espinosa-Moreno[1], Fernando García-García[1], Dae-Jin Lee[1,2], María J. Legarreta-Olabarrieta[3], Susana García-Gutiérrez[3], Naia Mas[4]

[1] Basque Center for Applied Mathematics (BCAM); Bilbao, Spain.
[2] IE University, School of Science and Technology; Madrid, Spain.
[3] Galdakao-Usansolo University Hospital, Research Unit; Galdakao, Spain.
[4] Galdakao-Usansolo University Hospital, Critical Care Unit; Galdakao, Spain.

E-mail for correspondence: `jcespinosa@bcamath.org`

Abstract: In this study, we propose a framework for analysing in-hospital patient data from electronic health records. We transform longitudinal sparse vital signs measures into cross-sectional data using descriptive statistics, imputing missing data and evaluating strongly associated variables with time until deterioration or favourable medical discharge. We employ competing risk and random survival forest techniques to predict the length of stay of patients and evaluate their performance to identify the most accurate model.

Keywords: Electronic health record; Competing risks; Variable selection model.

1 Introduction

The evaluation of health deterioration or favourable discharge of hospitalised patients, using vital signs as the main input, is usually based on electronic health records (EHR) which most of the time result in sparse data sets with some problems such as high rates of missing data. In this work, we aim to model the length of in-hospital stay (LoS) of each patient, making use of their vital signs with sex and age, taking as an objective variable the time until one of the two possible final states of each patient happens first. For this, we will employ competing risk models such as Fine and Gray's model, cause-specific Cox proportional hazard regression and random survival forest (in the same way that cause-specific Cox works) to model the time-to-event as a function of imputed cross-sectional data.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 Materials and methods

2.1 Data description and pre-processing

For each hospitalisation, we have the patient's sex and age, as well as longitudinal data along the hospitalisation for 7 vital signs: Temperature, systolic and diastolic pressure, heart and respiratory rates, oxygen saturation and neurological state. We summarise these longitudinal data with the following statistics: maximum, minimum, first observation, last observation, mean, standard deviation, average percentage change (avc) and average change per unit time (acpu), transforming the original variables into a cross-sectional higher dimensional space. We used the Multiple Imputation by Chained Equations (MICE) method for data imputation. In the Galdakao-Usansolo University Hospital (Basque Country, Spain), a total of 19.602 hospitalisations (lengths of stay at least 24 hours) were collected during the year 2019, of which 852 (4.35%) resulted in deterioration. These data correspond to 55.8% males and 44.2% females. Those data are split into train and testing data (70% and 30%, respectively), via stratified random sampling, to keep the proportion of events. Training data has 13722 hospitalisations with 597 (4.35%) that result in deterioration. Otherwise, the test has 5880 hospitalisations with 255 (4.33%) in deterioration.

2.2 Variable selection

To detect which variables are strongly associated with time until the final state, where possible events are deterioration and favourable discharge, we employ the LASSO Regularized Cox Regression (Simon *et al.,* 2009) and Best Subset Selection (Wen *et al.,* 2017) in CoxPH models. In LASSO, we obtain the best regularisation parameter λ by K-fold cross-validation (CV). In each one, LASSO and BeSS (Best Subset Selection), we define two models: (a) One using deterioration as an event and favourable discharge as censored data, where we obtained a set s_1 of variables; (b) one with deterioration as censored and favourable discharge as the event, where we obtain a set s_2 of variables. Finally, we define the definite set of variables as $s = s_1 \cup s_2$, which is a subset of the full set of variables.

2.3 Time-to-event models

Given that hospitalisations can result in two possible final states, deterioration or favourable discharge, we make use of competing risk models. The first model that we use is the Cause-Specific Cox (Austin *et al.,* 2016), where the hazard function denotes the instantaneous rate of occurrence of the k-th event in subjects who are currently event free and is calculated, for each possible event $D \in \{1, \cdots, K\}$, as described below.

$$\lambda_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t, D = k | T \geq t\right)}{\Delta t}, \tag{1}$$

where T is the random variable "baseline time until the occurrence of the event of interest" (such as death, failure, etc.), $t \in [0, \infty)$ and, in our case, $K = 2$. The second model, known as Fine and Gray (Austin *et al.,* 2016), also known as the sub-distribution hazard function, defines the instantaneous risk of failure from the kth event in subjects who have not yet experienced an event of type k (hazard function), as in Equation 2.

$$\lambda_k^{sd}(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t, D = k | T > t \ \cup (T < t \cap D \neq k)\right)}{\Delta t}. \quad (2)$$

The third model is the random survival forest (Ishwaran *et al.,* 2008), which estimates a set of survival trees that are grown using techniques such as bootstrap, splitting nodes based on feature selection and making an ensemble of all the trees to get one estimation of the survival function, based on the hazard function. Here, we adapted the survival random forest in the same way that cause-specific Cox, where we define the hazard function as in Equation 1. We make use of this methodology and make two random survival forests: one with deterioration as the event and favourable discharge as censored data, and vice-versa. We have then two models with different estimations of survival, one per cause and the hyper-parameters are tuned using out-of-sample error, avoiding the use of a validation data set. Finally, we use the Brier score, which is used to evaluate the accuracy of a predicted survival function at a given time t, for each possible final state. For a data set of N individuals, survival times T_i, covariables $\mathbf{X}_i$ and predicted survival function $\hat{S}(t)$, the Brier score is defined as $\text{BS}(t, S) = E(1_{T_i \geq t} - \hat{S}(t|\mathbf{X}_i))^2$, calculated for each final state.

3 Results and conclusions

Employing LASSO and BeSS, Table 1 summarizes the variables that are discarded to model the time to patient deterioration or discharge. None means that any of the summarised statistics were discarded for the method. We can see that the statistic more discarded was the first observation and the mean.

TABLE 1. Variables discarded by LASSO and BeSS methods.

Variable	Discarded by LASSO	Discarded by BeSS
Temperature	avc	first
Systolic pressure	first, mean	first
Diastolic pressure	avc	None
Heart rate	mean	avc
Respiratory rate	max, first, mean, sd, acpu	min, last, mean, acpu
Oxygen saturation	mean, acpu	None
Neurological state	None	None

We compare the Cause-Specific Cox (CSC), Fine and Gray (FG) and Random Survival Forest adapted as CSC (RSF-CS) concerning three models: the full model, which employs all the variables; a model with the variables obtained from LASSO Cox regression; finally, a model with the variables obtained from BeSS method. Then we calculated the Brier score for a LoS of 2,3,4 and 5 days (48, 72, 96 and 120 hours).

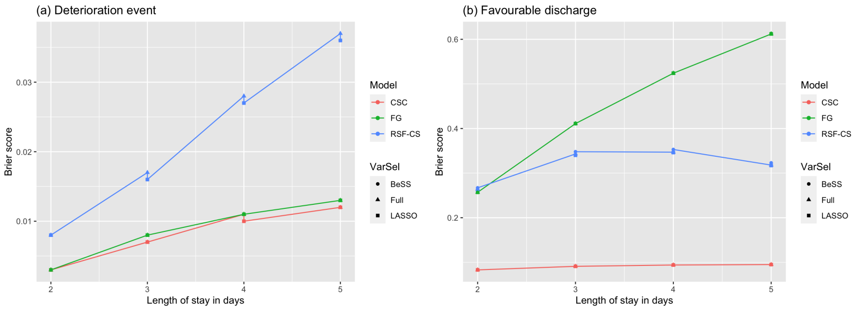

FIGURE 1. Brier score for the proposed models and variable selection methods.

As we observe in the last figure, for both states, we do not have a significant gain (in terms of Brier score) when we reduce the variables that are included in the models. Otherwise, between models, we can observe that cause-specific Cox was the best model to predict the time until both deterioration and favourable discharge. For deterioration, the Fine and Gray model has similar Brier scores to cause-specific and random survival forests had the worst performance. In another hand, for favourable discharge, the random survival forest presented the best performance than the Fine and Gray model, but they were far from cause-specific Cox.

References

Austin P.C., Lee D.S., Fine J.P. (2016). *Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation*, **133(6)**, 601 – 609.

Ishwaran H., Kogalur UB., Blackstone EH., Lauer MS. (2008). *Random survival forests. The Annals of Applied Statistics*, 841 – 860.

Simon N., Friedman J., Hastie T., Tibshirani R. (2009). *Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J Stat Softw*, **39(5)**, 1 – 13.

Wen C., Zhang A., Quan S. and Wang X. (2020). *BeSS: An R Package for Best Subset Selection in Linear, Logistic and Cox Proportional Hazards Models. J Stat Softw*, **94(4)**.