Competitive risk models in early warning systems for in-hospital deterioration: the role of missing data imputation

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

[1]{jcespinosa,fegarcia}@bcamath.org, Basque Center for Applied Mathematics (BCAM)
[2]daejin.lee@ie.edu, School of Science & Technology, IE University, Madrid, Spain
[3]{mariajose.legarretaolabarrieta,susana.garciagutierrez}@osakidetza.eus, Galdakao-Usansolo University Hospital, Research Unit
[4]naia.masbilbao@osakidetza.eus, Galdakao-Usansolo University Hospital, Critical Care Unit

Early Warning Systems (EWS) are useful and very important tools for evaluating the health deterioration of hospitalised patients, using vital signs (such as heart rate, temperature, etc.) as the main input, based on electronic health records (EHR) which most of the time result in sparse data sets with high rates of missing data. In this work, we aim to study the effect of different imputation techniques on time-to-event (survival) models.

For each case we have 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 central tendency, order and dispersion statistics: maximum, minimum, first observation, last observation, mean, standard deviation, average variance percentage and average derivative, transforming the original variables into a cross-sectional higher dimensional space, that still having missing data problems. Each hospitalisation has two possible final states: clinical deterioration or favourable discharge. Here, we model the time-to-event with competitive risk models taking into account the covariates.

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% of males and 44.2% of females. We are using a set of imputation methods, such as central tendency statistics (mean and mode), Multiple Imputation by Chained Equations (MICE), Non-Linear Principal Components Analysis (NLPCA) and Random Forest. We evaluate the performances of the imputation methods described before, via root mean square error and conclude the pros and cons of using each one in medical practice. Then, we use Fine and Gray's competitive risk models and the cause-specific Cox proportional hazard regression to model the time-to-event as a function of imputed summarised data. Finally, we evaluate these models employing the traditional and time-dependent area under the ROC curve, for horizon times of 24, 48, 72, 96 and 120 hospitalisation hours.

Keywords: Competing Risk models, Survival models, Data Imputation