The Effect of Missing Data when Predicting Readmission in Heart Failure Patients

Filip Plešinger¹, Zuzana Koscova¹, Eniko Vargova¹, Jan Pavlus¹, Radovan Smisek¹, Ivo Viscor¹, Veronika Bulkova²

¹Institute of Scientific Instruments of the CAS, Brno, Czechia
²Medical Data Transfer, s.r.o. Brno, Czechia

Background: The discharge of patients from hospital care is regulated by guidelines. Still, readmission of heart failure (HF) patients is a common issue, and several calculators have been published to predict it.

Aims: In this paper, we elaborate on how the prediction performance decreases when features become missing. We also elaborate on which features should a user include every time to reach acceptable prediction performance.

Method: We prepared a balanced dataset from HF patients in the MIMIC-III database (N=2,004) with 16 features. Using training data (80%) in a four-fold cross-validation manner, we evaluated all feature combinations (N=2¹⁶-1) and found the optimal feature set for the logistic regression model. We also evaluated feature presence in top-performing models (N=655) and identified essential features. Finally, we trained the resultant model using all training data and evaluated the effect of missing features (N=2⁸ combinations) using separate test data (20%).

Results: We identified five less important and three essential features (age, blood urea nitrogen, and systolic blood pressure). This led to a resultant model with eleven features. The hazard ratio (HR) using test data showed a value of 1.99 (95%CI 1.57-2.51) when all eleven features were present. It also showed an HR of 1.72 (95%CI 1.37-2.17) when only three essential features were present, and others were missing (i.e., replaced by zeros).

Figure: Feature presence (●) in 655 (1%) best performing models by mean test F1 score (four-fold cross-validation)