

*Table S1.* Differences in ROC (Panel A) and F statistic (Panel B) among the MLTs employed to predict in-hospital death. The plots report the confidence intervals for the difference in performances across resamples. The values below the diagonal of the tables represent the p-values of performance comparison across MLTs, while the numbers above the diagonal are the average differences in performances across metrics.

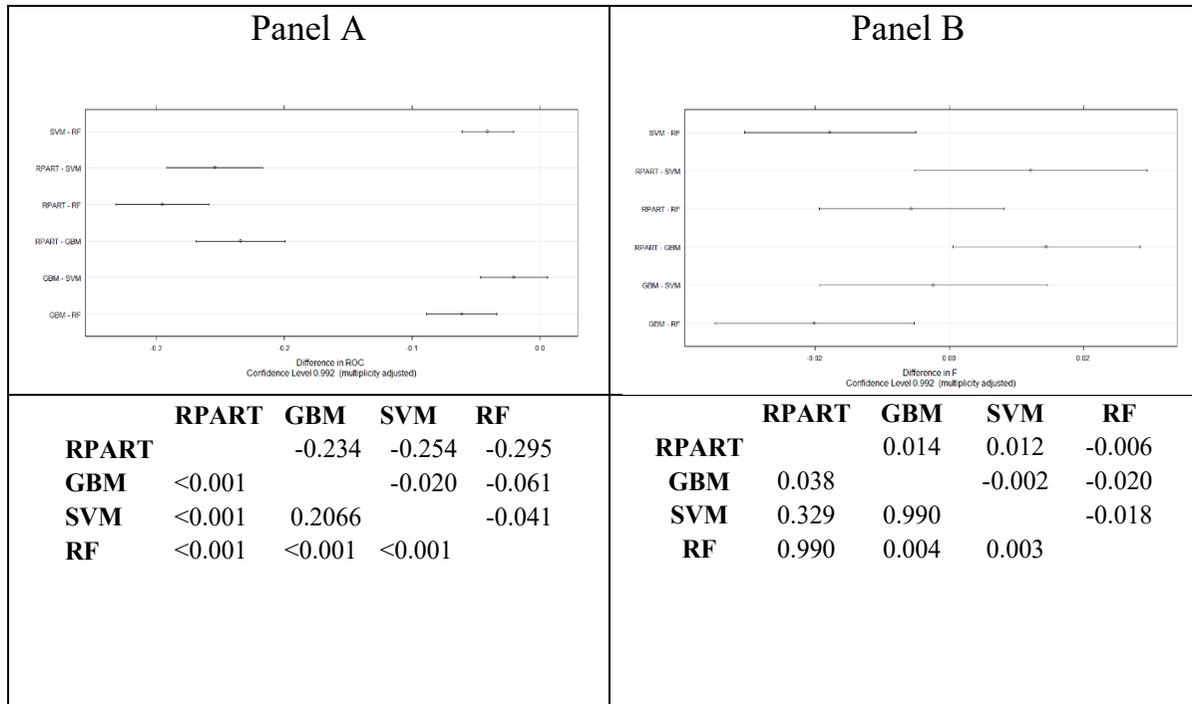
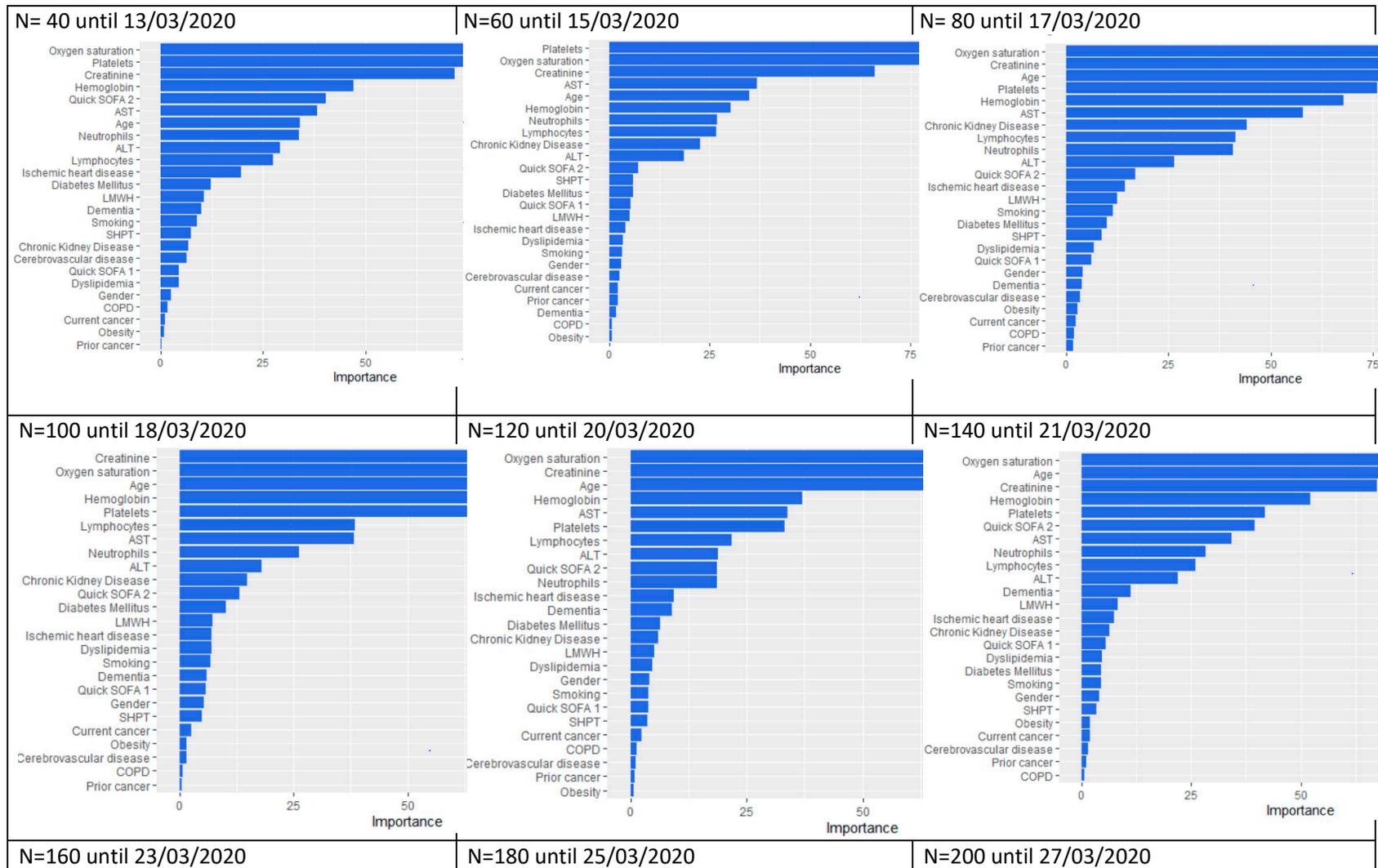
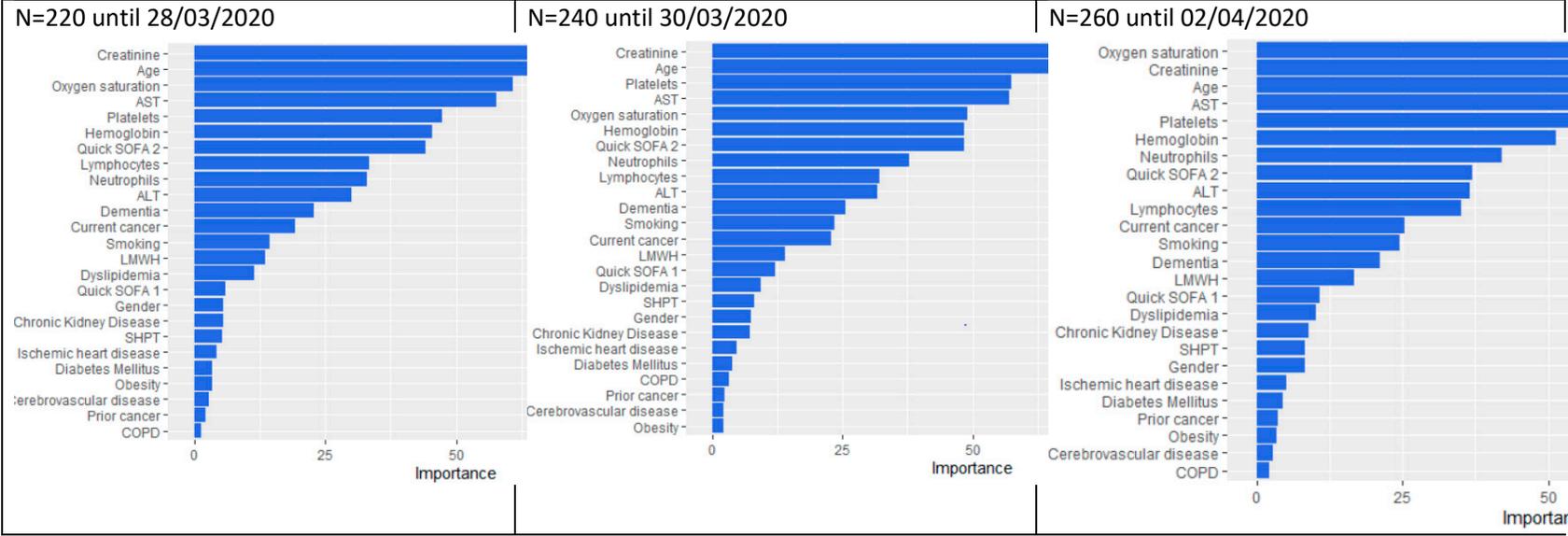
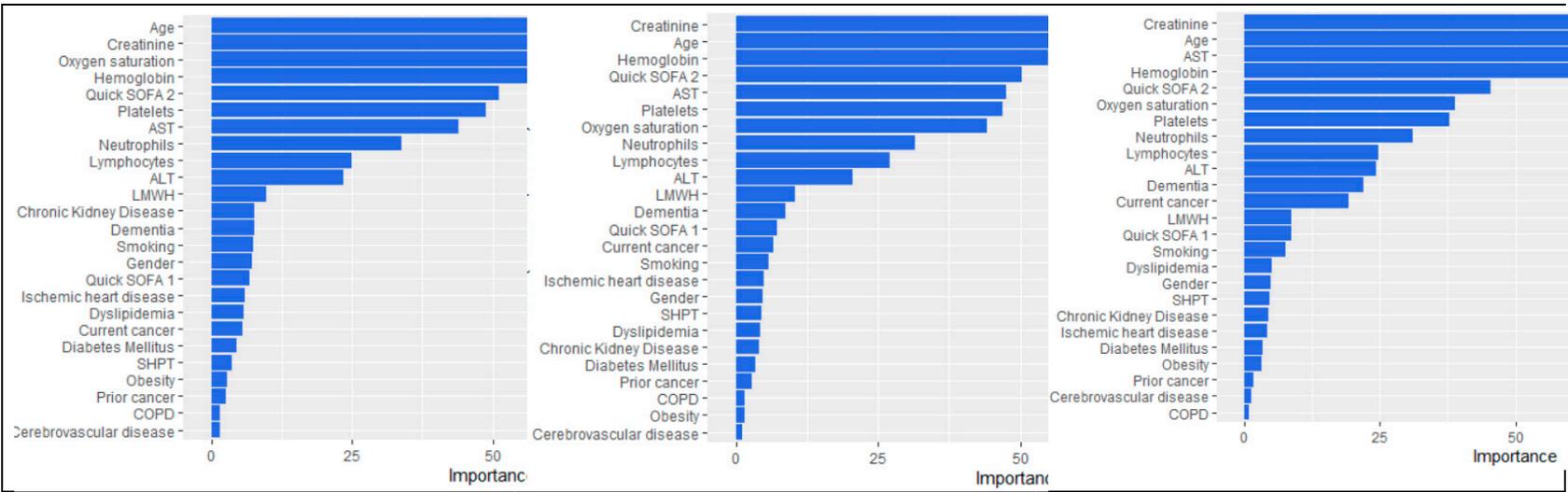


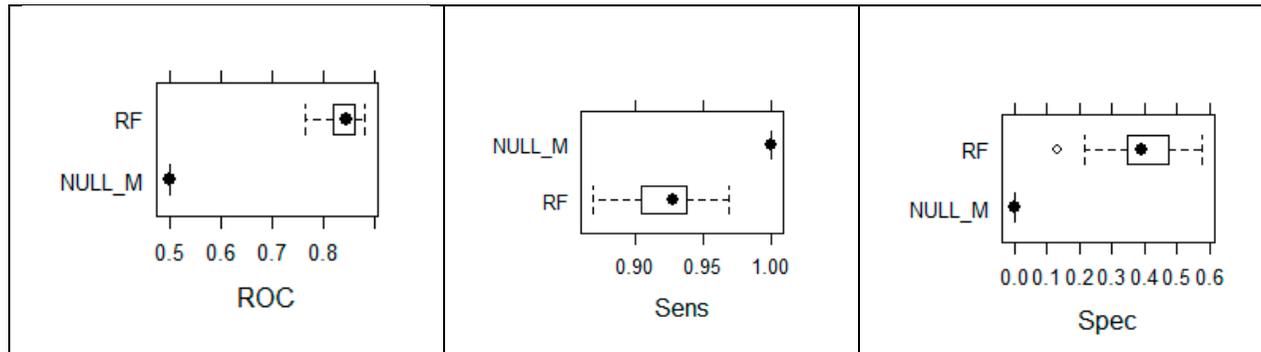
Table S2. Variable importance plots for each Random Forest developed.





*Figure S1.* To assess if class imbalance affected the results of the model with the best performance, i.e., the random forest, we tested the hypothesis of difference between the random forest and a null model. The null model represents the typical behavior of an MLT that does not properly handle unbalanced classes leading to biased predictions towards the majority class. The values of ROC, sensitivity, and specificity within the resampling procedures were reported for both the null model and the random forest. The difference between such measures was then tested using the approach proposed by Hothorn et al. [1] and Eugster et al. [2]. Significant differences between the null model and the random forest tool ( $P < 0.001$ ) were identified for the three measures of performance, demonstrating the random forest ability to deal with class imbalance.

The figure reports the ROC, sensitivity, and specificity measures of the null model and the random forest.



1. Hothorn, T.; Leisch, F.; Zeileis, A.; Hornik, K. The Design and Analysis of Benchmark Experiments. *Journal of Computational and Graphical Statistics* 2005, 14, 675–699.
2. Eugster, M.J.; Leisch, F. Exploratory Analysis of Benchmark Experiments an Interactive Approach. *Computational Statistics* 2011, 26, 699–710.