Supplementary Material S1

Machine Learning Insights

- Machine learning (ML) is an umbrella term that encompasses a variety of predictive algorithms (e.g.: neural networks, decision trees, random forests, k-means clustering, and others). Each algorithm presents peculiar strengths and pitfalls that render it more suitable to perform certain tasks.
- Random Forest (RF) is one of the most popular ensemble ML algorithms because of the easy hyperparameter tuning, the lower risk of bias compared with a simple decision tree, the elevated generalizability and accuracy, the ability to capture non-linear patterns in the data, and the applicability to different sizes of data. It works by creating an ensemble of different decision trees: majority "voting" of all the trees' outcomes is used to classify each patient's outcome. The RF algorithm can be used even when the number of features exceeds that of instances in the dataset and when most features are "noisy" (irrelevant to the outcome).
- Boruta is an all-relevant feature selection algorithm wrapped around the RF that allows to select and retain only the most predictive patient's variables (input features). Reducing input features has several advantages: first, decreasing the model's degrees of freedom decreases overfitting and improves generalizability; and second, promoting the most important patient's variables (or its linear combination) enhances the interpretability of the developed model.

Boruta's Initializing Parameters

- n_estimators: "auto,"
- verbose: 0,
- random state: 42,
- perc: 70,
- α: 0.05,
- two_step: "true,"
- max_iter: 100,

Random Forest Optimized Hyperparameters

- random state: 42,
- criterion: "gini,"
- n_estimators: 160,
- min_samples_split: 8,
- min_samples_leaf: 8,
- max_leaf_nodes: "None,"
- max_features: "log2,"
- max_depth: 4,
- bootstrap: "True"

