**Title**: Forecasting 1-year mortality following hospitalization using machine learning

**Authors**: Jason Dude, Tanner Reece, Keith Lohse

**Background:** Heart failure imposes a considerable health burden in the United States. For some patients, a hospital admission signals a shift in the disease trajectory, potentially indicating advancement to a higher-risk disease state and increased mortality. This highlights the necessity for precise predictive models to enhance the management and outcomes of affected individuals. Reliable prediction models estimating 1-year mortality post-hospitalization empower healthcare providers to devise personalized interventions, allocate resources efficiently, and enhance overall care quality.

**Methods:** The provided electronic health record data training set was partitioned further into train (85%) and validation (15%) sets using random sampling. Missing data were handled by 1) converting numeric variables with high missing rates to categorical using tertile splits plus an unknown level (i.e., low, medium, high, unknown) and 2) imputing variables with low missing rates using multiple imputation by chained equations (MICE). All numeric features were normalized prior to modeling, and features with high skewness were log-transformed. Machine learning algorithms were implemented using 10-fold cross-validation with hyperparameter tuning and grid search and evaluated using area under the curve (AUC): logistic regression (LR), random forest (RF), support vector machine (SVM), gradient boosted decision tree (GBDT). Variable importance plots from RF models were used to identify salient features. Finally, using an ensemble approach, probability predictions from top-performing models were combined and used as inputs into a logistic regression model. Several models were evaluated using the validation set, and the best model was re-fit on the combined train and validation sets and used for test set predictions.

**Results**: The train and validation sets had similar class proportions (~74% in both sets) after partitioning. Numeric variables with >15% missing values (N=27) were split into tertiles, while the MICE algorithm imputed missing values on all remaining numeric variables (N=41). AUC was similar across all tested algorithms (LR: .702, RF: 0.703, SVM: .687, GBDT: .692). Features deemed important by the RF model were selected (N=8) and implemented with logistic regression (LR-sub) using all two-way interactions, resulting in marginally lower performance compared to full models (AUC: .681). Probability predictions from the full LR (LR-full) and RF models were combined and used as inputs into another LR model, resulting in an AUC of .714, which was the highest obtained AUC during training. The LR-full and LR-sub were the two models selected for validation set predictions, and the LR-full was selected for test predictions after obtaining better performance during validation (LR-full: .743, LR-sub: .705).

**Conclusion:** Despite our best efforts to employ advanced machine learning algorithms and modeling techniques, simple logistic regression models performed equally well when training. In fact, the variability of AUC was often greater between various hyperparameter selections of the same model than between models. One potential explanation for this is that the outcome (1-year mortality) may be approximately linearly separable, and we thus do not obtain benefits from implementing advanced models with complex decision boundaries. Although we did generate an ensemble model with a marginally better AUC, we chose logistic regression for validation assessment and subsequently test predictions for several reasons: logistic regression is a simple algorithm that does not require hyperparameter tuning/selection, the loss function is convex, allowing for rapid model training and low computational cost, and the algorithm has a straightforward interpretation of model results and coefficients. Logistic regression may be a simple, effective method for quantifying risk of 1-year mortality following hospitalization.