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

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

**Background:** Heart failure poses a catastrophic risk for an individual and a considerable burden for the health system in the United States. For some patients, a hospital admission signals a shift in the disease trajectory, potentially indicating disease progression and increased mortality. Clinically meaningful predictive models would be of tremendous value to both providers and patients, leading to more efficient disease management and improved outcomes for affected individuals. The goal of this study was to identify reliable models predicting 1-year mortality post-hospitalization with a clinically meaningful degree of accuracy, helping empower healthcare providers to devise personalized interventions, allocate resources efficiently, and enhance quality of care.

**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 missingness into categorical factors using tertile splits plus an unknown level (i.e., low, medium, high, unknown) and 2) imputing variables with low missingness using multiple imputation by chained equations. 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 then evaluated using the area under the curve (AUC) from signal detection theory. Specifically, we used: logistic regression (LR), random forest (RF), support vector machine (SVM), and gradient boosted decision trees (GBDT). Finally, we also tested an ensemble approach, using probability predictions from top-performing models as inputs into a second-level logistic regression model.

**Results**: The train and validation sets had similar class proportions after partitioning (~74% in both sets). 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). Additionally, estimated class probabilities from the full LR (LR-full) and RF models were combined and used as inputs into an ensemble LR model, resulting in an AUC of .714, only marginally better than simpler LR models. The LR-full and LR-sub were 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, simpler logistic regression models performed equally in the training data. 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 have linear associations with predictors, and we thus do not obtain benefits from implementing models with more complex decision boundaries. Although we did generate an ensemble model with a marginally better AUC, we chose logistic regression for validation and test predictions for several reasons: logistic regression is a simpler algorithm that does not require hyperparameter tuning/selection, the log-loss function is convex, allowing for quick optimization and low computational cost, and the algorithm has a straightforward interpretation of model results and coefficients. Logistic regression may be an effective method for quantifying risk of 1-year mortality following hospitalization.