Novel Leveraging of Perioperative Data for Early Diagnosis of Heart Failure: A Machine Learning Approach

Michael R. Mathis, MD[1]; Hyeon Joo, MS[1]; Milo C. Engoren, MD[1]; Brahmajee K. Nallamothu, MD, MBA[2]; Michael Burns, MD, PhD[1]; Kayvan Najarian, PhD[3]; Sachin Kheterpal, MD, MBA[1]

Departments of Anesthesiology,[1] Cardiology,[2] and Computational Bioinformatics,[3] Michigan Medicine, Ann Arbor, MI

Background. Heart failure (HF) is a condition imposing significant national healthcare burden. Despite advances in treatment, there remain limitations to reliable and inexpensive early diagnostic measures, as impeded by the complex, multifactorial nature of the disease. Many patients undergo routine surgery in the US, and perioperatively, a wealth of health data is collected including cardiovascular-specific tests. Perioperative clinicians are well-positioned to identify patients with HF in early stages, and initiate cardiologist referral to commence key therapies. However, clinician synthesis of health data is challenged by limited resources for full, timely reviews. To improve accuracy and reduce costs, machine learning (ML) techniques are a promising option to sift through data and improve identification of patients with undiagnosed HF.

In this study, we use ML techniques to understand patterns which may serve to diagnose HF with reduced ejection fraction (HFrEF) - a subtype of HF amenable to detection via retrospective EHR data - in preclinical stages utilizing data collected in a perioperative setting. We hypothesize that patients with undiagnosed HFrEF can be identified via ML methods.

Methods. This is an observational study at our academic tertiary care medical center. Following ethical approval, we extracted data from the Multicenter Perioperative Outcomes Group and our institution (Epic Systems, Verona, WI).

Study Population. We examined adult (>40 years) patients undergoing non-cardiac surgical procedures from 2010-2016. We excluded cardiac procedures, as *undiagnosed* early-stage HFrEF would be unexpected due to extensive preoperative cardiac evaluation. We similarly excluded patients with preoperative ventilation, inotropes, mechanical circulatory support, previous heart/lung transplantation, or moribund (American Society of Anesthesiologists Physical Status Class 5, 6, or postoperative mortality). We excluded minor procedures (e.g., office-based) with limited preoperative evaluation.

HFrEF Phenotype. We classified patients into three phenotypes: (i) *healthy controls*; (ii) *known HFrEF* (patients with pre-existing HFrEF); and (iii) *undiagnosed HFrEF* (patients without a preoperative HF diagnosis but with HFrEF diagnosed within 2 years postoperatively). A clinician expert (MRM) verified *undiagnosed HFrEF* phenotype patients via manual review, excluding/reclassifying patients as necessary. Importantly, this excluded cases for which HFrEF was secondary to a perioperative triggering event (e.g. cardiac arrest, sepsis, cardiotoxic chemotherapeutic agents following tumor resection, etc.), or any other event not associated with HFrEF natural disease progression. Following review, *known HFrEF* patients were excluded. The target output (binary primary outcome) was *undiagnosed HFrEF.*

Model Features. A total of 261 features were collected and included patient demographics, anthropometrics, comorbidities, preoperative vitals, laboratory values, medications, testing observations, and surgical / anesthetic characteristics. Features were binned into categorical variables, including missing/unknown, and inspected for validity and data leakage.

Modelling. We partitioned the data into training/validation (80%) and test (20%) sets. As we expected a skewed target output distribution due to low HFrEF frequency, we used balanced class weights inversely proportional to class frequencies. Within the training/validation set, we trained ML algorithms using 5-fold cross validation to select parameters maximizing receiver operating characteristic area under the curve (AUC). We used standard approaches (L2 regularization, lower range of tree depth) to mitigate overfitting. After performing cross-validation and grid search to optimize parameters, we retrained the models using all data from training/validation set. The selected best models were tested on the test data set.

Results. Our final analytic dataset included 68,387 *healthy controls* and 285 patients with *undiagnosed HFrEF*. Table 1 describes model performance. Extreme gradient boosting was the best model (AUC 0.847; 95% CI 0.781-0.913) on cross-validation, but logistic regression was the best on the test set (AUC 0.850). All test results were within 95% CI of expected performance.

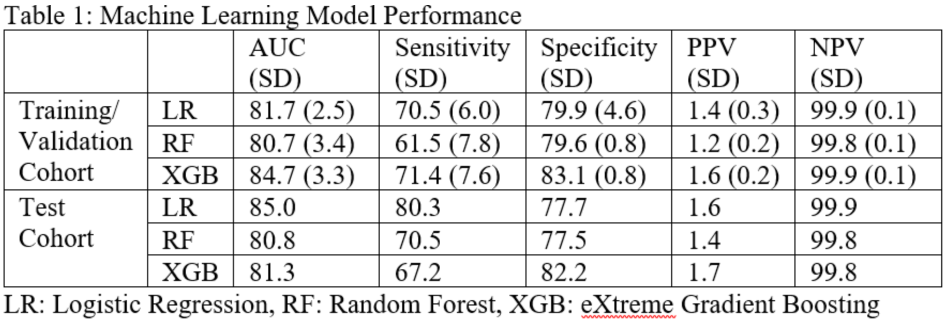

Table 1: Machine Learning Model Performance

		AUC (SD)	Sensitivity (SD)	Specificity (SD)	PPV (SD)	NPV (SD)
Training/ Validation Cohort	LR	81.7 (2.5)	70.5 (6.0)	79.9 (4.6)	1.4 (0.3)	99.9 (0.1)
	RF	80.7 (3.4)	61.5 (7.8)	79.6 (0.8)	1.2 (0.2)	99.8 (0.1)
	XGB	84.7 (3.3)	71.4 (7.6)	83.1 (0.8)	1.6 (0.2)	99.9 (0.1)
Test Cohort	LR	85.0	80.3	77.7	1.6	99.9
	RF	80.8	70.5	77.5	1.4	99.8
	XGB	81.3	67.2	82.2	1.7	99.8

LR: Logistic Regression, RF: Random Forest, XGB: eXtreme Gradient Boosting

Conclusion. Given the low frequency of the target output, traditional ML models predicted *undiagnosed HFrEF* with good performance, but limited positive predictive value. However, confirmatory testing (e.g. B-type natriuretic peptide level, echocardiography) may be able to identify such patients amenable to a prospective cardiology referral for evaluation prior to surgery, with clinically useful performance; this remains a future direction of our current research.