Prediction of Adverse Events in Patients Undergoing Major Cardiovascular Procedures

Bobak J. Mortazavi, Nihar Desai, Jing Zhang, Andreas Coppi, Fred Warner, Harlan M. Krumholz, and Sahand Negahban

***Abstract*—Electronic health records (EHR) provide opportunities to leverage vast arrays of data to help prevent adverse events, improve patient outcomes, and reduce hospital costs. This paper develops a postoperative complications prediction system by extracting data from the EHR and creating features. The analytic engine then provides model accuracy, calibration, feature ranking, and personalized feature responses. This allows clinicians to interpret the likelihood of an adverse event occurring, general causes for these events, and the contributing factors for each specific patient. The patient cohort considered was 5214 patients in Yale-New Haven Hospital undergoing major cardiovascular procedures. Cohort-specific models predicted the likelihood of postoperative respiratory failure and infection, and achieved an area under the receiver operating characteristic curve of 0.81 for respiratory failure and 0.83 for infection.**

***Index Terms*— Cardiology, electronic health records, machine learning, outcomes, Prediction.**

I. Introduction

THE early prediction of potential adverse events in patients has been a primary focus of outcomes research and quality improvement efforts in patient care for heart failure [1], readmissions [2], and a variety of other outcomes [3]. These efforts have focused improving patient care in a wide variety of fields, including in early detection of severe events in infants [4], respiratory complications in surgical patients [5], and blood transfusions in cardiac surgery patients [6], by understanding factors leading to conditions like costly readmissions [7], septic shock [8], and unplanned transfers to the intensive care unit [9]. These targeted models for care can help identify patient risk factors and predictors [10], [11] as well as potentially address costs of care [12], [13].

One major area of research focuses on surgical complications [14], [15] and understanding the risk factors involved [16], [17] to predict outcomes [18], [19]. In particular, understanding complications such as the risk of infection [8] and respiratory failure [17], [20], and other outcomes post-cardiac procedures is a particular area of focus for care [21], [22] and cost [13]. Electronic health records (EHR) have been viewed as an increasingly useful source of data for such outcomes research across varying patient cohorts and outcomes predictions [3], [23], [24]. Research on EHR data has ranged from better patient history representation [25], [26] to subtyping patient backgrounds [27] for better precision medicine applications and personalized risk predictions [7], [10], [28]. Recent efforts have aimed at developing patient condition scores to be used for outcomes modeling cases [29], [30]. However, with varying EHR systems and a variety of admissions criteria, it is important to understand the data available for outcomes modeling in specific patient populations.

This work will develop a system for identifying patients undergoing cardiovascular procedures at risk for postoperative complications using preoperative EHR data. The procedures considered are coronary artery bypass grafting (CABG), percutaneous coronary intervention (PCI), and implantable cardioverter defibrillators (ICD), and will model postoperative respiratory failure and infection. This system will focus on the extraction of all data from the time of admission to either the start of the procedure or the end of the first twenty-four hours of admission, whichever comes first. This time period has been identified by the Yale-New Haven Hospital as useful for understanding patient risk factors and determining potential interventions. The data will be extracted for use in a machine learning framework to predict patient risk as well as identify the top factors for that risk. Patients and clinicians can use this risk to make better informed decisions on treatment plans with better knowledge about the risk.

II. Related Works

A. Electronic Health Record Models

Several works have focused on using EHR data to predict outcomes. In [10], authors investigated the use of EHR data to predict readmissions in heart failure patients. Authors extracted patient information (including age, gender, marital status), specific visit information (date, duration, inpatient or outpatient visit, and source of admission), as well as visit information broken up into categories of patient history, labs, medications, and the attending physicians. Using a lasso technique to select the most relevant binary features for the statistical model, au-

Manuscript received August 15, 2016; revised December 21, 2017 and February 13, 2016; accepted February 17, 2017. Date of publication March 1, 2017; date of current version November 3, 2017. *(Harlan M. Krumholz and Sahand Negahban contributed equally to this work.)*

B. J. Mortazavi and S. Negahban are with the Center for Outcomes Research and Evaluation, Yale School of Medicine, and the Department of Statistics, Yale University, New Haven, CT 06510 USA (e-mail: bobak.mortazavi@yale.edu; sahand.negahban@yale.edu).

N. Desai, J. Zhang, A. Coppi, F. Warner, and H. M. Krumholz are with the Center for Outcomes Research and Evaluation, Yale School of Medicine, Yale University, New Haven, CT 06510 USA (e-mail: nihar.desai@yale.edu; jing.zhang@yale.edu; andreas.coppi@yale.edu; frederick.warner@yale.edu; harlan.krumholz@yale.edu).

Digital Object Identifier 10.1109/JBHI.2017.2675340

thors were able to achieve an area under the Receiver Operating Characteristic (ROC) curve (AUC) of 0.71 and demonstrate potential cost savings. This work will similarly examine the details of EHR data. It will investigate the use of a lasso technique for feature selection in building a logistic regression model. Given the wide array of data types, it will also employ other methods that are better suited for higher dimensional and varied data types.

Work in [8] developed a real-time risk score for septic shock using EHR data. Using the MIMIC dataset available on PhysioNet, authors extracted suspicion of infection via ICD-9 codes, used a multiple imputation approach for missing information or unknown/censored events, and developed an advanced model based upon Cox proportional hazards and lasso regularization for estimating risk. This paper's work aims to approach prediction problems similarly, outlining the data extraction and developing a method to generate predictions; however, since this work aims to evaluate predictions at a specific time, the methods used are varied for this purpose, to leverage the cross-sectional data since continuous data as in MIMIC is usually restricted to intensive care units.

B. Rothman Index

The Rothman Index, by PeraHealth, is a patient condition score based upon EHR data [29]. This score is built off of 26 variables extracted from medical record data for patients during hospital admissions. In particular, the variables are broken up into vital signs, laboratory tests, cardiac rhythm information, and a variety of nursing assessments that are converted into met/unmet variables [29]. The design of the score was to help quantify patient condition based upon data generated by nurses during admissions.

There are two predictive models developed using the Rothman Index as the primary feature [31], [32]. Work in [31] developed a predictive model for unplanned 30-day readmissions using the Rothman Index at discharge, age, gender, insurance type, and service type (medical or surgical). A logistic regression model built from this data had an AUC of 0.73 and the Rothman Index score was shown to be correlated to higher odds of readmission, with an AUC of only 0.68 when the Rothman Index was removed. However, by removing the Rothman Index, the model is left with only the service type for the clinical information. This work will also consider the effectiveness of the Rothman Index as a way to summarize EHR data in a meaningful manner, but will compare it with use of other clinical data extracted from the medical records.

Work in [32] used the Rothman Index to predict unplanned surgical intensive care unit readmissions, by evaluating the range of Rothman Index scores generated during stays and correlating them to the transfers. However, while evaluating the importance of first and last Rothman Index scores, no predictive models were built to consider the effects of a variety of Rothman Index scores throughout the patient encounter to predict adverse events. This work will develop predictive models for post-surgical outcomes through a variety of modeling techniques based upon increased Rothman Index data availability and increased EHR data availability.

III. Method

This section details the personalized predictions of postoperative complications in cardiovascular procedure patients. It also covers the extraction of data from the EPIC electronic health record system [33] used by Yale-New Haven Hospital (Y-NHH). The cohort consisted of patients admitted to the Heart and Vascular Center (HVC) for cardiac procedures, with a primary principal procedure code for CABG, PCI, or ICD. This study used all data available in the EHR from February, 2013 (the go-live date for EPIC at Y-NHH) through September, 2015. As prior data were stored on a different EHR system, all visits from this date forward were considered first visits. Methods considered for this work considered only data upon patient presentation at admission and collected from then forward. As a results, no outpatient data, including emergency room visit data that led to the admission was included, except for the source of admission, to understand the transfer-in status of the patient. For each patient, if multiple visits occurred, only the first visit was considered, though the lack of prior visit data lends the methods developed to repeated use. Outcomes of respiratory failure and infection were defined by the Quality Variation Indicators^{TM} (QVI) developed by Yale-New Haven Hospital to identify those patients with adverse events developed postoperation, which result in poor patient outcomes and extensive cost to the medical system [13], [34]. 111 patients passed away after the procedure, with only 46 being within 48 hours of procedure. This study was approved by the Yale HIC (# 1506015993).

A. Data Source

Data were extracted for each admission. Each visit's dataset consisted of data from admission time to either 24 hours or the start of patient's first procedure, whichever came first; this period of time was believed to be long enough to gather clinically relevant information on the patients to provide an understanding of patient risk prior to the procedure that resulted in the adverse event. Further, this aligned with clinical rounds typically happening every morning and procedures often happening soon after admission. The desired goal, therefore, was to create a dataset and system that would serve as a balance between early enough for appropriate decision making and late enough for considering a wide array of data. The following categories of information were gathered:

1) Patient Information: Included features such as age, gender, insurance, and admission information.
2) Patient History: Included information such as the patient problem list and admission diagnosis codes (ICD-9).
3) Visit Information: Included primary principal procedure information, admission time, and attending staff information.
4) Medical Information: Included medications prescribed, laboratory results, and patient vitals, including temperature, pulse oxygenation, systolic blood pressure, diastolic blood pressure, respiratory rate, and heart rate.
5) Rothman Index: Rothman Index scores.

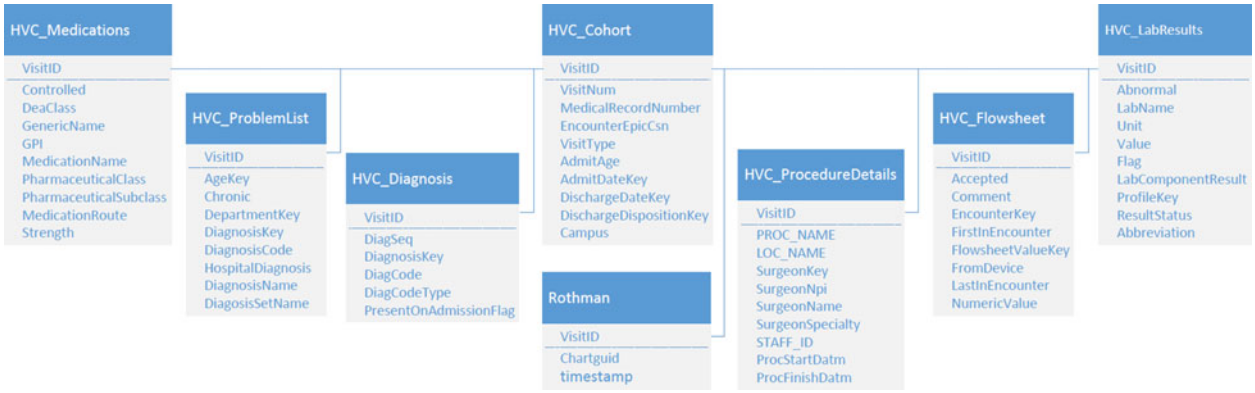

Fig. 1. Query and gathering of patient data from the EHR.

TABLE I
DATABASE TABLE DESCRIPTIONS

Table Name	Number of Columns	Number of Rows	Description
Cohort	181	5214	Patient information, demographics, admission, accounting, and outcomes
Diagnosis	8	89 988	Diagnosis information in ICD codes
Flowsheet	26	27 906 192	Patient data, vitals, exams, misc. notes
Lab Results	38	2502 649	Lab results, including abnormal flags, and reference values
Medications	52	380 544	Medications prescribed during admission, including name, dose, and class
Procedure Details	11	11 559	Procedure details and attending physicians
Problem List	20	230 278	Prior history diagnosis information

TABLE II
PATIENT POPULATION AND EVENT RATE

Primary Procedure	Respiratory Failure	Infection
CABG ($n = 1025$)	64 (6.2%)	29(2.8%)
PCI ($n = 2539$)	53 (2.1%)	20 (0.8%)
ICD ($n = 1650$)	41 (2.5%)	25 (1.5%)
Total ($n = 5214$)	158 (3.0%)	74 (1.4%)

The data were extracted from the EHR data tables as in Fig. 1, where each VisitID in the patient cohort table had a one to many relationship with entries in each of the other tables of the database. The data were organized in seven tables (plus a Rothman Index Scores table), listed in Table I. These tables were joined from back-end tables storing data from the front-end of EPIC. The Cohort table contained patient information, including the admission source (e.g., self-referral, transfer from another hospital, transfer from another unit, physician referral), insurance information (e.g. medicare, private insurance, etc.), and personal information (e.g. age, gender, race if provided). The patient population included 1025 CABG patients, 2539 PCI patients, and 1650 ICD patients. Table II shows the event rates of respiratory failure and infection. Despite the low event rates, these patients were adversely harmed and attributed a significant cost to the hospital [13]. No unstructured text was extracted from the EHR. In particular, any data that would require natural language processing was left for future work. The data extracted were structured data organized in the back-end data warehouse for the EHR system, allowing for quick manipulation of fields for feature extraction.

B. Feature Extraction

Once the appropriate data were extracted from the EHR, it needed to be converted into a format suitable for use in machine learning analytics. Much of the information was stored in a one-to-many format needing manipulation. For example, in Fig. 1, medication information was stored in a fashion where a single VisitID might consist of multiple rows in the database, where the medication name and pharmaceutical class fields contained each prescribed medication information.

All categorical variables were created into distinct binary yes/no variables for each factor. For example, the problem list and diagnosis information for each visit were converted into a series of binary yes/no variables for each individual ICD-9 code, lab results had a yes/no for lab conducted and results available. The yes/no variable allows the machine learning algorithm to understand if the remaining extracted lab variables, namely numeric results and alert flags (based upon stored reference values), were missing values or reported results from a conducted lab.

The flowsheet table contained many of the structured vital sign information for each patient. As vitals may have been taken multiple times between admission and procedure start time, a time-series was generated for each variable, as was for the Rothman Index. Features for the length of the time-series, as well as the mean, standard deviation, minimum, and maximum were created as well. Since this created variable-length time-series, each patient's first and last readings were saved, the windowed features calculated, then additional readings were dropped, rather than determine an appropriate imputation. More complex methods might find spurious patterns in the specific

readings if improperly imputed. Time-series data were represented by first reading, last reading, number of readings, mean, minimum, maximum, and standard deviation. For laboratory readings, only the last laboratory reading was considered, due to the sparse nature.

1) Grouping of Variables: The extraction of the dataset resulted originally in 14 353 variables per patient. This set of features included 1764 prior history variables and diagnosis codes, 8328 variables for laboratory information, 1942 variables for medication information, and 2319 variables for patient admission information. Thus, some dimension reduction became necessary. The machine learning methods used (discussed below in Section III-D) were selected because of their abilities to select a sparse set of features from a high-dimensional set such as this. Preliminary dimension reduction, however, could be done manually, by changing the specificity of the features created. Taking guidance from medical expertise as well as national registries such as the National Cardiovascular Data Registry (NCDR) [35], features were merged whenever clinically appropriate. For example, the 1577 binary variables from medication/dosage information were reduced to 295 variables of medication counts via the use of pharmaceutical class. More explicitly, rather than have a variable for each dosage of aspirin given (125 mg vs. 165 mg), these were combined into a variable that includes just aspirin, and this was combined further to the pharmaceutical class of all the medications. Similar techniques were applicable to the insurance information, race information, and laboratory information. Prior history variables were grouped together when known chronic condition flags were met. This reduced medication to 295 variables, grouped prior history variables, laboratory, and others as well, by eliminating those with no variance, and this reduction of variables resulted in a final set of 9828.

2) Missing Variables: The potential for missing data after extraction is an important issue in EHR datasets. Data might be missing for a variety of reasons, from the patient chose not to disclose race information, to laboratory results that were normal did not set the flag variables, and are dependent upon the implementation strategy and completeness in filling out the interactive forms and transmitting that data to the backend databases. In many cases, binary indicator variables can easily be imputed with a 0/no if not present for a given visit (i.e., 0 indicates either missing or not prescribed medication, 1 is a definitive prescription of a medication). For any missing variable that could not similarly be coded, such as numeric vital sign information as well as Rothman Index, it was determined that missing data should be imputed with the mean value, since a 0 Rothman Index score, for example, would indicate a severely ill patient. This imputation occurred after the training sets and testing sets were created, using only the training means, so that no knowledge of the testing data was included in this calculation.

3) Normalization: After the dataset is created, it was z-scored (centered and scaled) by subtracting the feature mean and dividing by the feature standard deviation. If the feature standard deviation was 0 the feature was removed entirely.

C. Validation

A cross-validation framework was setup to analyze the effectiveness of the proposed methods. Many clinical papers often use a single 80/20 random split to create their training and testing datasets [1], [2]. This work used a five-fold stratified cross-validation in order to create similar 80/20 splits and maintain the observed event rate in each fold. The imputation steps as well as the normalization, indicated above, were carried out after the folds were created, with the training means being used to impute both the training set and the testing set alike, and the training means and standard deviations being used to normalize the training set and the testing set. The system layout for validation is represented in Fig. 2.

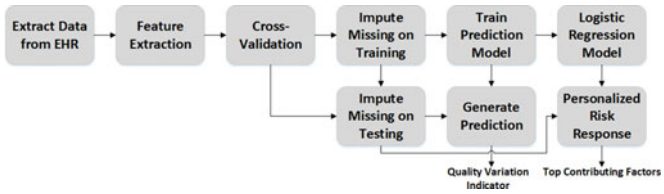

Fig. 2. System diagram for data analytic engine.

D. Data Analytic Engine

Once the training set was created, it was passed to three different modeling techniques. Those techniques were logistic regression with lasso regularization (a form of generalized linear model), random forest, and gradient descent boosting. The analysis was carried out in R, with the glmnet being the chosen implementation for the logistic regression and generalized linear model approach (hereafter GLM) [36], randomForest for the random forest algorithm (hereafter RF) [37], and xgboost or eXtreme Gradient Boosting as the implementation of a gradient descent boosting method chosen (hereafter XGB) [38] respectively. These techniques were selected due to their ability to select a sparse set of features while training, to avoid overfitting, and further reduce the dimensionality of the problem, where applicable. Further, GLM is commonly used in clinical practice and outcomes research, linking to similarity in related works, while RF and XGB are particularly good at dealing with data of mixed types such as these, by setting differing thresholds in each particular decision tree. Further, as these last two are non-linear methods, they might provide stronger results than linear methods commonly used in clinical outcomes research.

1) Hyperparameter Tuning: For GLM, an internal cross-validation on the training data was run in order to tune the algorithm hyperparameters, with the AUC being the optimized measure. Sample weights were provided, where the weight for each adverse event example was the ratio of dataset size to number of adverse outcomes (the inverse of the event rate). The default parameters were selected for RF, and XGB was tuned using a grid-search for the number of iterations (100 to 1000 in 100 step-size increments) and the maximum depth of each tree (5 to 10) in an internal cross-validation.

E. Prediction

Models were trained on the entire dataset as well as created by patient cohort and outcomes splits. Once trained, each algorithm generated a response for the test set. This response was a generated probability of a postoperative complication, rather than a

strict label output. From this, an ROC curve plot allowed calculation of an AUC. AUCs are often reported in clinical prediction models [1], due to the measure being unaffected by class imbalance [39]. However, to understand how such models would be used prospectively, more information should be presented regarding the predictive accuracy. After the models and AUCs were generated, an optimal threshold probability was selected to generate the classification labels. The threshold selected was that which maximized the F-score. A further discussion of the optimal point is left for Section VI. From this classification, the true positives, true negatives, false positives, and false negatives were calculated and from that an F-score. Finally, a further metric was calculated regarding the precision of the top 20 predictions, to see if all the true positives are captured in the riskiest patients predicted as a numeric measure for how well calibrated the algorithm is. The 20 were selected based upon the total number of adverse events in each sub-group, knowing that a subset of these would exist in each fold, and to evaluate if creating a larger interval would account for all the true positives or not. This value can be altered to highest deciles of risk, quartiles, and the definitions should be created in consulation with the clinical professionals involved to understand their desires of evaluating 'high-risk' patients. For all the measures, the mean and 95% confidence intervals were calculated. Calibration plots were also created for the best models generated.

F. Personalized Risk Factors

The ability to interpret model predictions is highly desirable for clinicians, and to potentially help determine risk factors resulting in the prediction and potentially helping determine interventions or actions that might prevent the postoperative complication. While the models provided the selected global features, feature importance was extended to provide patient-specific results. Namely, GLM provided a vector of $\overrightarrow{\beta} = \langle \beta_1, \beta_2, \dots \rangle$ coefficients for each parameter, which provide the global feature importance and where the length of the vector is equal to the number of features (and a large number are 0 for non-selected features). For every test patient $\tilde{x} = \langle x_1, x_2, \dots \rangle$ the component-wise multiplication of the two vectors results in a feature-contribution vector $\overrightarrow{feat} = \langle \beta_1 \times x_1, \beta_2 \times x_2, \dots \rangle$ whose components are then summed together by GLM for the resulting prediction. Sorting these components then provided the clinicians with the top contributing factors of risk for each individual patient.

IV. Results

A. Test Framework

The analysis presented in Sections III-D, III-E, and III-F was run on the five-fold cross-validation dataset. As a reminder, all data were used from the admission time until either the first procedure start or 24 hours, whichever came first. All time-series based features used considered all available data in this window. In order to evaluate the effectiveness of all the features generated from the EHR, and to compare against methods previously generated using the Rothman Index [31], [32], the following four Rothman tests [31], [32], as well as two configurations with the data extracted in this paper, were created, over the same extraction window as the remaining data:

1) Rothman Index test using patient demographics, history, insurance, and the earliest Rothman Index - hearafter 'eRI'.
2) Rothman Index test using eRI as well as mean, standard deviation, minimum, and maximum - hereafter windowed 'eRI'.
3) Rothman Index test using patient demographics, history, insurance, and the latest Rothman Index - hearafter 'lastRI'.
4) Rothman Index test using lastRI as well as mean, standard deviation, minimum, and maximum - hereafter 'windowed lastRI'.
5) EHR dataset - all extracted features without the Rothman Index features - hereafter 'EHR-RI'.
6) Complete EHR Dataset - all extracted features including the Rothman Index features - hereafter 'EHR'.

B. Single Model Tests

The first tests designed were run in order to validate the effectiveness of separating patients by procedures as well as outcome. Table III shows the best single model AUC and the model type that generated it for each test type and patient cohort. Further, the final two columns show the mean F-score and mean precision of the top 20 generated risk scores. While the top 20 precision is likely increased due to the larger number of cases to train and test on, the lower AUC indicates that only the highest risk is well identified. Indeed, the similar F-scores show that, even with high precision, recall is affected, and that only the highest risk patients are well identified. It became clear that some prediction results were strengthened by specifying the patient population, likely due to the different risks associated with each procedure type. The remainder of the tests evaluated the hypothesis that multiple models should be developed for the prediction of postoperative complications for the patient procedures due to the patient heterogeneity in each case.

C. Respiratory Failure

Models were created separately for CABG patients, PCI patients, and ICD patients to predict respiratory failure. The results for each can be found in Tables IV, V, and VI, respectively. For each test case, GLM, RF, and XGB models were created, with the strongest model's mean AUC and mean F-score over cross-validation presented. The mean precision of the top 20 predicted risks are also presented to present an interpretation of model calibration independent of the cutoff threshold selected to generate the F-score. This means that, for the top 20 patients when sorted by outputted risk score, the precision was then calculated on these patients only.

1) CABG Patients: Note that for CABG patients, in Table IV, using the windowed information of the Rothman Index provided a higher AUC (mean AUCs 0.59 and 0.58 for windowed eRI and windowed lastRI respectively). Using the last Rothman Index helped provide higher F-score for an F-score of 0.22 for windowed lastRI. In all cases, the use of EHR data provided higher AUC (0.60 for both cases) but a slightly lower

TABLE III
BEST MEAN AUC (AND MODEL THAT GENERATED IT) FOR PREDICTING POSTOPERATIVE COMPLICATIONS

Test Configuration	Rothman				EHR-RI	EHR	Mean F-score	Mean Top 20 Precision
	eRI, windowed eRI, lastRI, windowed lastRI							
All Patients	0.62 (GLM)	0.62 (GLM)	0.62 (GLM)	0.62 (GLM)	0.66 (GLM)	0.66 (GLM)	0.36	0.68
CABG Patients	0.59 (RF)	0.59 (GLM)	0.59 (RF)	0.60 (RF)	0.61 (RF)	0.61 (RF)	0.39	0.43
PCI Patients	0.62 (GLM)	0.64 (GLM)	0.65 (GLM)	0.65 (GLM)	0.65 (RF)	0.67 (GLM)	0.30	0.37
ICD Patients	0.64 (GLM)	0.66 (GLM)	0.66 (GLM)	0.65 (GLM)	0.66 (RF)	0.67 (RF)	0.34	0.36

TABLE IV
BEST MEAN AUC (95% CONFIDENCE INTERVAL (CI), MODEL) FOR PREDICTING RESPIRATORY FAILURE IN CABG PATIENTS

Test Configuration	Mean AUC (95% CI - Model)	Mean F-score	Mean Top 20 Precision
eRI	0.57 (0.51–0.64, RF)	0.22 (0.15–0.29)	0.00
windowed eRI	0.58 (0.50–0.66, RF)	0.20 (0.14–0.26)	0.00
lastRI	0.59 (0.48–0.70, RF)	0.18 (0.11–0.24)	0.00
windowed lastRI	0.58 (0.50–0.66, RF)	0.22 (0.13–0.30)	0.00
EHR-RI	0.60 (0.53–0.68, XGB)	0.18 (0.11–0.25)	0.07
EHR	0.60 (0.56–0.65, GLM)	0.20 (0.12–0.27)	0.07

TABLE V
BEST MEAN AUC (95% CONFIDENCE INTERVAL (CI), MODEL) FOR PREDICTING RESPIRATORY FAILURE IN PCI PATIENTS

Test Configuration	Mean AUC (95% CI - Model)	Mean F-score	Mean Top 20 Precision
eRI	0.62 (0.53–0.71, GLM)	0.12 (0.01–0.22)	0.04
windowed eRI	0.63 (0.45–0.81, XGB)	0.15 (0.05–0.24)	0.07
lastRI	0.66 (0.59–0.73, GLM)	0.19 (0.10–0.28)	0.11
windowed lastRI	0.67 (0.48–0.85, XGB)	0.17 (0.07–0.27)	0.08
EHR-RI	0.80 (0.70–0.90, RF)	0.24 (0.11–0.37)	0.00
EHR	0.81 (0.70–0.92, RF)	0.25 (0.12–0.37)	0.00

TABLE VI
BEST MEAN AUC (95% CONFIDENCE INTERVAL (CI), MODEL) FOR PREDICTING RESPIRATORY FAILURE IN ICD PATIENTS

Test Configuration	Mean AUC (95% CI - Model)	Mean F-score	Mean Top 20 Precision
eRI	0.75 (0.67–0.83, GLM)	0.20 (0.17–0.24)	0.00
windowed eRI	0.76 (0.76–0.85, GLM)	0.24 (0.20–0.27)	0.00
lastRI	0.73 (0.58–0.87, GLM)	0.27 (0.23–0.32)	0.00
windowed lastRI	0.76 (0.66–0.86, GLM)	0.24 (0.21–0.28)	0.00
EHR-RI	0.79 (0.65–0.94, RF)	0.30 (0.14–0.46)	0.00
EHR	0.78 (0.64–0.93, RF)	0.27 (0.15–0.40)	0.00

F-score (0.18 and 0.20 for EHR-RI and EHR respectively). The EHR-RI and EHR had a more defined high-risk group with the top 20 measure of 0.07 in both cases. While the best CABG model was GLM, the similar AUC across each data configuration and each method indicates that linear models performed sufficiently well, and further investigation is necessary to understand why RF and XGB did not provide better results. For the model with the highest F-score, the EHR model, the features selected in each fold can be found in the supplement material. These features were sorted by largest absolute coefficient in GLM, largest mean decrease in accuracy as ranked by RF, and by model information gain, in XGB, and the top features are listed here:

1) *Fold 1:* Respiration Rate, Prior History: Hypovolemia, Lab: Blood Urea Nitrogen (BUN) is High, Primary Diagnosis: Coronary Atherosclerosis of Native Coronary Artery.
2) *Fold 2:* Prior History: Hypovolemia, Lab: Prothrombin Time is Abnormal, Lab: MCH is unspecified.
3) *Fold 3:* Earliest Respiration Rate, Lab: Albumin, Prior History: Hypovolemia, Lab: Albumin.
4) *Fold 4:* Earliest Heart Rate, Prior History: Hypovolemia, Lab: PO2 Arterial, Med: Serotonin-2 Antagonist, Patient Demographics: Race - Other, Primary Diagnosis: Coronary Atherosclerosis of Native Coronary Artery.
5) *Fold 5:* Prior History: Other or Unspecified Hyperlipidemia, Primary Diagnosis: Coronary Atherosclerosis of Native Coronary Artery.

As described in Section III-B, the flags and thresholds are predetermined by the laboratory and defined within the table in EPIC.

2) PCI Patients: All models for PCI patients, presented in Table V, were able to better predict respiratory failure than in CABG patients or in ICD patients. Similar to CABG patients, using the windowed information of the Rothman Index provided a higher AUC than the single measure (mean AUCs 0.63 and 0.67 for windowed eRI and windowed lastRI respectively). Using the last Rothman Index helped provide higher F-score for an F-score of 0.19 for lastRI. In all cases, the use of EHR data provided significantly higher AUC measurements from both the single model for PCI patients (0.67) and any of the Rothman Index test cases, with an AUC of 0.80 for EHR-RI and 0.81 for EHR. Similarly, the F-score for these two cases were higher as well, at 0.24 and 0.25 respectively. However, none of the cases performed well in the top 20 precision measure. For the model with the highest F-score, the EHR model, the top features are listed here:

1) *Fold 1:* Prior History: Acute Respiratory Failure, Med: Analgesics Narcotic- Anesthetic Adjunct Agents, Lab: ECG - P Axis, Lab: Glucose Meter is Low, Prior History: Acute Myocardial Infarction of Inferolateral Wall Episode of Care Unspecified.
2) *Fold 2:* Med: Analgesics Narcotic- Anesthetic Adjunct Agents, Med: IV Solutions - Dextrose Water, Prior History: Acute Respiratory Failure, Admit Source: Self Referral, Lab: MCHC.

3) *Fold 3:* Med: Analgesics Narcotic- Anesthetic Adjunct Agents, Prior History: Acute Respiratory Failure, Lab: ECG - P Axis, Lab: CO2, Lab: Glucose Meter is Low.
4) *Fold 4:* Prior History: Acute Respiratory Failure, Lab: CO2, Prior History: Cardiogenic Shock, Lab: MCHC, LAB: Bun to Creatinine Ratio.
5) *Fold 5:* Med: Analgesics Narcotic- Anesthetic Adjunct Agents, Med: IV Solutions - Dextrose Water, Lab: Glucose Meter is Low, Lab; B-type Natriuretic Peptide ProBNP is Abnormal, Lab: Bands Present is Abnormal.

3) ICD Patients: ICD patient respiratory failure predictions, presented in Table VI, were improved over the single model AUC of 0.67 from Table III. The Rothman Index models performed better than the single model case, as well, with the windowed eRI and windowed lastRI each achieving the higher AUC of 0.76. Using the last Rothman Index score improved the F-score of the models to 0.27. The EHR-RI and EHR models performed the best, with the RF models achieving an AUC of 0.79 and 0.78 respectively and F-scores of 0.30 and 0.27 respectively. For the model with the highest F-score, the EHR-RI model, the top features are listed here:

1) *Fold 1:* Prior History: Acute Respiratory Failure, Primary Diagnosis: Acute on Chronic Systolic (Congestive) Heart Failure, Primary Diagnosis: Combined Systolic and Diastolic Heart Failure - Acute on Chronic, Admit Source: Self Referral, Med: Sodium-Saline Preparations.
2) *Fold 2:* Primary Diagnosis: Systolic Heart Failure - Acute on Chronic, Prior History: Acute Respiratory Failure, Admit Source: Physician or Clinical Referral, Admit Source: Self Referral, Lab: Glucose Meter.
3) *Fold 3:* Prior History: Acute Respiratory Failure, Primary Diagnosis: Systolic Heart Failure - Acute on Chronic, Admit Source: Self Referral, Primary Diagnosis: Combined Systolic and Diastolic Heart Failure - Acute on Chronic, Lab: Lactate.
4) *Fold 4:* Admit Source: Self Referral, Admit Source: Emergency, Primary Diagnosis: Systolic Heart Failure - Acute on Chronic, Prior History: Intermediate Coronary Syndrome - Unstable Angina, Lab: ECG T Wave Axis.
5) *Fold 5:* Prior History: Acute Respiratory Failure, Primary Diagnosis: Systolic Heart Failure - Acute on Chronic, Admit Source: Self Referral, Primary Diagnosis: Combined Systolic and Diastolic Heart Failure - Acute on Chronic, Lab: Potassium is High Panic.

D. Infection

Results for the models developed for infection are presented in Table VII for CABG patients, Table VIII for PCI patients, and Table IX for ICD patients, respectively.

1) CABG Patients: Models on CABG patients, in Table VII, using the windowed information of the Rothman Index did not provide the higher AUC, which was achieved by eRI at 0.67. Windowed eRI had the same AUC, however, provided a tighter confidence interval as well as provided a higher F-score at 0.41. The additional EHR data did not provide any improved AUC or F-score, and had a reduced top 20 precision of 0.00 down from 0.12. For the model with the highest F-score, the EHR model, the top features are listed here:

1) *Fold 1:* Prior History: Congestive Heart Failure - Unspecified, Present On Admission: Respiratory Failure, Present on Admission: Sepsis, Admit Source: Self Referral, Lab: INR.
2) *Fold 2:* Prior History: Congestive Heart Failure - Unspecified, Lab: Anion Gap, Med: Solvents, Present On Admission: Respiratory Failure, Med: Heparin.
3) *Fold 3:* Prior History: Unspecified Glaucoma, Primary Diagnosis: Unspecified Septicemia, Present On Admission: Respiratory Failure, Med: Sodium-Saline Preparations, Lab: Partial Thromboplastin Time is High Panic.
4) *Fold 4:* Prior History: Congestive Heart Failure - Unspecified, Present On Admission: Respiratory Failure, Lab: PH UA is Abnormal, Lab RDW, Lab: Amorphous is Abnormal.

TABLE VII
BEST MEAN AUC (95% CONFIDENCE INTERVAL (CI), MODEL) FOR PREDICTING INFECTION IN CABG PATIENTS

Test Configuration	Mean AUC (95% CI - Model)	Mean F-score	Top 20 Precision
eRI	0.67 (0.50–0.85, GLM)	0.32 (0.14–0.50)	0.12
windowed eRI	0.67 (0.54–0.80, RF)	0.41 (0.24–0.58)	0.00
lastRI	0.65 (0.50–0.80, GLM)	0.32 (0.22–0.43)	0.11
windowed lastRI	0.65 (0.52–0.79, RF)	0.40 (0.23–0.58)	0.00
EHR-RI	0.66 (0.54–0.77, RF)	0.29 (0.21–0.38)	0.00
EHR	0.67 (0.53–0.81, RF)	0.29 (0.19–0.39)	0.00

TABLE VIII
BEST MEAN AUC (95% CONFIDENCE INTERVAL (CI), MODEL) FOR PREDICTING INFECTION IN PCI PATIENTS

Test Configuration	Mean AUC (95% CI - Model)	Mean F-score	Top 20 Precision
eRI	0.72 (0.54–0.89, XGB)	0.10 (0.00–0.20)	0.03
windowed eRI	0.71 (0.54–0.88, XGB)	0.11 (−0.05–0.27)	0.01
lastRI	0.64 (0.43–0.84, XGB)	0.10 (−0.01–0.27)	0.02
windowed lastRI	0.61 (0.54–0.88, XGB)	0.13 (−0.06–0.21)	0.02
EHR-RI	0.81 (0.66–0.95, XGB)	0.12 (0.04–0.21)	0.03
EHR	0.83 (0.72–0.93, XGB)	0.14 (0.04–0.23)	0.04

TABLE IX
BEST MEAN AUC (95% CONFIDENCE INTERVAL (CI), MODEL) FOR PREDICTING INFECTION IN ICD PATIENTS

Test Configuration	Mean AUC (95% CI - Model)	Mean F-score	Top 20 Precision
eRI	0.56 (0.46–0.67, GLM)	0.06 (0.01–0.11)	0.02
windowed eRI	0.68 (0.52–0.85, GLM)	0.17 (0.06–0.28)	0.06
lastRI	0.64 (0.50–0.77, GLM)	0.11 (0.05–0.18)	0.03
windowed lastRI	0.67 (0.53–0.81, GLM)	0.16 (0.10–0.21)	0.00
EHR-RI	0.78 (0.65–0.91, RF)	0.16 (0.10–0.23)	0.00
EHR	0.78 (0.64–0.92, RF)	0.18 (0.11–0.25)	0.00

5) *Fold 5:* Prior History: Congestive Heart Failure - Unspecified, Med: Sodium-Saline Preparations, Present On Admission: Respiratory Failure, Admit Source: Self-Referral, Present on Admission: Severe Sepsis.

2) PCI Patients: Models on PCI patients, presented in Table VIII, were able to better predict infection than in CABG patients or ICD patients. Similarly to CABG patients, using the earliest Rothman Index provided a higher AUC (0.72). In all cases, the use of EHR data provided significantly higher measurements from both the single model for PCI patients (0.67) and any of the Rothman Index test cases, with an AUC of 0.81 for EHR-RI and 0.83 for EHR, as well as an F-score of 0.12 and 0.14 respectively. The top 20 precision measurements were higher for PCI patients as well, as a measure of identifying high risk patients. For the model with the highest F-score, the EHR model, the top features are listed here:

1) *Fold 1:* Admission: Age, Med:Adrenergic Vasopressor Agents, Lab: Enterovirus by RT-PCR Stool is Abnormal, Lab: POC Activated Clotting Time is Abnormal, Med: Anithypertensives.
2) *Fold 2:* Admission: Age, Lab: Albumin (EP) Urine Random is Abnormal, Med: Antivirals, Lab: Activated Protein C Resistance is Abnormal, Lab; Cortisol Plasma is Abnormal.
3) *Fold 3:* Admission: Age, Lab: Fibrinogen Level, Lab: Vitamin D 25 Hydroxy is Abnormal, Lab: HCV Quantitative Log is Abnormal, Prior Coverage is Other.
4) *Fold 4:* Admission: Age, Prior History: Acute Respiratory Failure, Lab: POC Appearance UA is Abnormal, Lab: Fluid Culture, Lab: POC Leukocytes UA is Abnormal.
5) *Fold 5:* Admission: Age, Lab: Antibody Identification is Abnormal, Lab: Protein Creatinine Ratio Urine Random is Abnormal, Lab: Cocaine Screen Urine, Med: Folic Acid.

3) ICD Patients: ICD patient infection predictions, presented in Table IX, were improved over the single model AUC of 0.67 from Table III. The Rothman Index models performed better than the single model case, as well, with the windowed eRI and windowed lastRI achieving AUCs of 0.68 and 0.67. Windowed eRI had the highest F-score of 0.17. The EHR-RI and EHR models performed the best, with the RF models achieving an AUC of 0.78 and 0.79 respectively and F-scores of 0.16 and 0.18 respectively. No model had top 20 precision. For the model with the highest F-score, the EHR model, the top features are listed here:

1) *Fold 1:* Primary Diagnosis: Combined Systolic and Diastolic Heart Failure - Acute on Chronic, Lab: Absolute Lymphocyte Count, Lab: Glucose Meter, Med: Sodium-Saline Preparations, Lab: International Normalization Ratio (POC).
2) *Fold 2:* Primary Diagnosis: Combined Systolic and Diastolic Heart Failure - Acute on Chronic, Lab: Bilirubin Total, Lab: Absolute Lymphocyte Count, Admit Source: Self Referral, Lab: Glucose Meter.
3) *Fold 3:* Primary Diagnosis: Systolic Heart Failure - Acute on Chronic, Admit Source: Self Referral, Primary Diagnosis: Combined Systolic and Diastolic Heart Failure - Acute on Chronic, Med: Sodium-Saline Preparations, Lab: ECQ QT Interval.
4) *Fold 4:* Primary Diagnosis: Systolic Heart Failure - Acute on Chronic, Admit Source: Self Referral, Primary Diagnosis: Combined Systolic and Diastolic Heart Failure - Acute on Chronic, Admit Source: Physician or Clinic Referral, Med: Sodium-Saline Preparations.
5) *Fold 5:* Primary Diagnosis: Systolic Heart Failure - Acute on Chronic, Primary Diagnosis: Combined Systolic and Diastolic Heart Failure - Acute on Chronic, Lab: International Normalization Ratio POC, Admit Source: Self Referral, Admit Source: Physician or Clinic Referral.

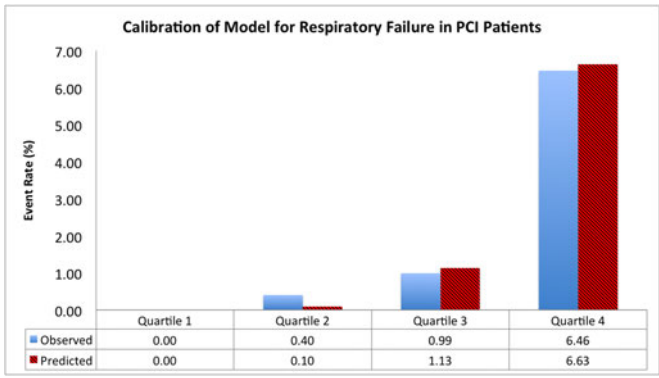

Fig. 3. PCI patient observed respiratory failure rate per quartile of risk.

E. Calibration and Personalized Risk

Understanding the factors behind the risk and outcome predicted is equally important to an accurate model. Thus, the system provided model calibration plots to better interpret patient risk. One such plot, for the model generating respiratory failure risk for PCI patients, is shown in Fig. 3. The calibration plot was created by sorting the probabilities generated by the model for the outcome into quartiles, then comparing the observed rate of respiratory failure to the mean risk for all predictions in each quartile. As shown in Fig. 3, quartile 1 has no observed respiratory failure predictions, thus, the high F-score of 0.25 and AUC of 0.81, despite the 0.00 Top 20 precision measure. This indicated that, while the model was able to generate a high risk group (quartile 4), the stratification within that group had room for improvement. Such calibration plots allow clinicians to better interpret the accuracy measurements generated by the models to understand underlying risk.

Further, along with the generated model accuracy, predictions, and calibration plots, the important features that generate the risk for a given patient were important in determining a cause and potential intervention. While each method provided a global list of important features, how each feature contributes to an individual's total risk score should be understood. Thus, the system generates an identification of which risk quartile the patient lies within, as well as the personalized response to the GLM model, as detailed in Section III-F. As an illustrative example, the GLM for the PCI respiratory failure, which achieved a mean AUC of 0.76 used the following features:

1) Lab 1 - Blood Urea Nitrogen is High - $\beta = 0.0910$.
2) Lab 2 - Anion Gap is High - $\beta = 0.1124$.
3) Med 1 - Anti-Hyperlipidemic - HMG COA Reductase Inhibitors Given - $\beta = -0.0142$.
4) Primary Diagnosis - Coronary atherosclerosis of native coronary artery - $\beta = -0.2751$.

Consider the following two patient $\overrightarrow{\text{feat}}$ vectors. The patient risk for patient x_1 was 0.61 while the patient risk for patient x_2 was 0.62. Both patients did, indeed, have respiratory failure, as correctly indicated by the model. However for x_1, $\overrightarrow{\text{feat}(x_1)} = \langle 0.273, 0.337, -0.014, 0\rangle$ while for x_2, $\overrightarrow{\text{feat}(x_2)} = \langle 0.273, 0.337, -0.028, 0\rangle$. This specific level of information illustrated the top contributors to the patient's specific risk score were, which could be extremely important in cases where the models might select hundreds of variables. In this particular case, the second patient had had more medication than the first, slightly increasing the predicted risk. The validation of the usefulness of this aspect of the system is left for further discussion in Section VI.

V. Discussion

A. Single Model Results

The results showed an interesting distribution of strengths and areas of necessary improvement. Having all patients together confounded the results, achieving low AUCs despite the methods employed and high top 20 precision. The added data did not appear to help for most patients. Thus, such settings were only ideal for identifying those at highest risk. Table III shows that evaluating each group individually lead to a better understanding of strengths and weaknesses. In particular, PCI and ICD patients improved over the all patients model, while CABG patients were reduced. Further work is necessary to understand if those individual CABG patients were better predicted by the all patients model, but it is likely that they were similarly missed there. Thus, separating models into individual ones for each patient group achieved greater success, enabling more specific results in future interventions. The system used the best available model knowing the particular patient. Understanding how this might change throughout the course of an intervention is left for Section VI.

B. Cohort-Specific Features and Results

For the respiratory failure and infection models, significant improvement was seen in the PCI patients and ICD patients. These models saw significant improvement by separating out the patient cohorts as well as incorporating the spectrum of EHR data selected. In these cases, the Rothman Index tests, with fewer variables, were well modeled by GLM, while RF and XGB provided the higher accuracy when the significantly wider array of variables were provided. In many cases, the EHR-RI and EHR models performed similarly. The Rothman Index provided some added value, but in all cases, the extension of the datasets to the EHR data provided the largest basis for improvement. As more features were added to the models, and the complexity increased, the non-linear, non-parametric methods were better suited to finding higher-dimensional patterns for prediction. This became quite apparent when looking at the top features selected for each model in each fold. The GLM models, best in CABG patients, selected mostly binary variables. In contrast, the RF and XGB models often chose continuous variables, and a spread of medication information, laboratory results, as well as prior history and patient presentation information. The reference value flags were often selected as well, which aligns thinking with clinical interpretability. Of note was that the top selected features for XGB were a majority of numeric laboratory results, rather than the flag values of the labs selected by RF and GLM. Further, the present on admission flags along with laboratory values for these tree-based methods may have allowed for the removal of a number of false positives, thus improving AUC and F-score (improved recall) but not top 20 precision.

The numeric results for AUC, F-score and top 20 also aligned with calibration results. In particular, the improved AUC values indicated a better opportunity for the models to discriminate patients. With the low AUCs in CABG, all following results were similarly low, because an effective threshold delineating adverse outcomes and healthy outcomes was not clear. The lower F-scores, with the improved AUCs, were a function of the event rate. The low score indicated that the recall (sensitivity) was high but the precision was low. So while the threshold for determining clearly healthy outcomes was well-established, the mix of true positive predictions and false positive predictions is still an area for further investigation. This was also demonstrated in the top 20 precision and the calibration results. The right-skewed calibration results indicated that the adverse outcomes were mostly in the highest quartile of risk. However, with the low top 20 precision, these patients were not the highest risk. An expansion of the binary outcomes to multiple classes, with tiered understandings of the postoperative period, might be necessary to understand these false positive patients and why they are predicted differently than the large number of correctly identified true negative patients. This may also be because of other events that are not currently recorded or considered adverse outcomes in this study. This is left for future considerations.

VI. Future Work

A number of future steps remain to validate the effectiveness of such a system. First, and foremost, is to continue collecting new patients, but it is important to also evaluate further machine learning methods in comparison, such as neural networks and support vector machines. In addition, the method by which we clean the data and impute missingness should be further explored, to understand, especially for time-series data, prior and future readings to better interpolate missing values. Further, the personalized risk factors are focused specifically on the logistic regression with lasso regularization, due to its familiarity in clinical literature [1], [2] as well as clear, linear interpretation. However, as the best models are achieved by random forests and gradient descent boosting, a more advanced way of understanding personalized feature effects in these settings should be developed. As a focus of this work was the ability for clinicians to interpret and potentially alter intervention strategies based

upon predicted risks, two forms of validation need to be carried out. First do the personalized risk factors teach the clinicians anything about prospectively treating patients or ordering new laboratory exams? Second, would these interventions reduce adverse events? The decision boundary considered for generating the F-score needs further evaluation including how much data is necessary to generate an accurate result and how early can this prediction be made in future time-based models.

Many clinical model papers present an AUC as a measure of the model's effectiveness of identifying patients with adverse events from those without; however, they do not tell clinicians how to prospectively identify patients at risk. A model can have strong calibration and still potentially have a low F-score, simply because of the number of observed events in the lower quartiles being well predicted. A selected cutoff threshold must consider the balance of true positives, true negatives, false positives, and false negatives and the costs associated with each, considering costs of alarm fatigue and treatments on false positives. Alternatively, the costs of a false negative might greatly outweigh a cost of a false positive. Once this information, is considered, a better optimal decision boundary could be calculated. Finally, as amount and variety of collected data grows, models can begin to consider multiple visits and outpatient visits. This includes data from emergency room visits that lead to admissions, as well as understanding re-admissions and risks associated from multiple in-hospital stays. This wider array of data can include other assessment scores besides the Rothman Index, including the Goldman Multifactorial, ASA physical status classification, Euroscore, and National Cardiovascular Data Registry models.

VII. Conclusion

This work developed a system for identifying patients undergoing major cardiovascular procedures at the Yale-New Haven Hospital at risk for postoperative respiratory failure or infection, two costly outcomes as identified by the hospital. This system tackles the challenges of extracting data from a production-level electronic health record provided by EPIC [33] and the tasks necessary in manipulating data for use in machine learning analytic tools. Further, after developing models to predict postoperative complications using preoperative data, the system generated interpretable measures of risk to help identify the risk category of the patient, as well as the contributing features to risk in order to better provide clinicians with information that might help prevent such adverse events, providing a framework for more advanced clinical decision support systems in future studies.

Acknowledgment

The authors would like to thank Yale-New Haven Hospital's S. Allegretto, K. Churchwell, M. Donini, R. Gibson, A. Green, B. McCloskey, K. Pont, J. Rimar, and C. Torre, Jr., for the problem description as well as assisting in data access, extraction, and dictionary building.

References

[1] K. Rahimi *et al.*, "Risk prediction in patients with heart failure: A systematic review and analysis," *JACC: Heart Failure*, vol. 2, no. 5, pp. 440–446, 2014.

[2] J. S. Ross *et al.*, "Statistical models and patient predictors of readmission for heart failure: A systematic review," *Archives Internal Med.*, vol. 168, no. 13, pp. 1371–1386, 2008.

[3] B. B. Dean, J. Lam, J. L. Natoli, Q. Butler, D. Aguilar, and R. J. Nordyke, "Review: Use of electronic medical records for health outcomes research a literature review," *Med. Care Res. Rev.*, vol. 66, no. 6, pp. 611–638, 2009.

[4] S. Saria, A. K. Rajani, J. Gould, D. Koller, and A. A. Penn, "Integration of early physiological responses predicts later illness severity in preterm infants," *Sci. Translational Med.*, vol. 2, no. 48, 2010, Art. no. 48ra65.

[5] J. P. Fischer, A. M. Wes, J. D. Wink, J. A. Nelson, B. M. Braslow, and S. J. Kovach, "Analysis of risk factors, morbidity, and cost associated with respiratory complications following abdominal wall reconstruction," *Plastic Reconstructive Surgery*, vol. 133, no. 1, pp. 147–156, 2014.

[6] G. J. Murphy, B. C. Reeves, C. A. Rogers, S. I. Rizvi, L. Culliford, and G. D. Angelini, "Increased mortality, postoperative morbidity, and cost after red blood cell transfusion in patients having cardiac surgery," *Circulation*, vol. 116, no. 22, pp. 2544–2552, 2007.

[7] R. Amarasingham *et al.*, "An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data," *Med. Care*, vol. 48, no. 11, pp. 981–988, 2010.

[8] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (TREWScore) for septic shock," *Sci. Translational Medi.*, vol. 7, no. 299, 2015, Art. no. 299ra122.

[9] J. A. Rubano, J. A. Vosswinkel, J. E. McCormack, E. C. Huang, M. J. Shapiro, and R. S. Jawa, "Unplanned intensive care unit admission following trauma," *J. Critical Care*, vol. 33, pp. 174–179, 2016.

[10] M. Bayati *et al.*, "Data-driven decisions for reducing readmissions for heart failure: General methodology and case study," *PloS One*, vol. 9, no. 10, 2014, Art. no. e109264.

[11] A. Visser, B. Geboers, D. J. Gouma, J. C. Goslings, and D. T. Ubbink, "Predictors of surgical complications: A systematic review," *Surgery*, vol. 158, no. 1, pp. 58–65, 2015.

[12] M. Bayati, S. Bhaskar, and A. Montanari, "A low-cost method for multiple disease prediction," in *AMIA Annu. Symp. Proc.*, vol. 2015. American Medical Informatics Association, 2015, p. 329.

[13] "Strata partners with yale new haven health system to reduce cost by improving quality," [Online]. Available: http://www.stratadecision.com/our-company/newsroom/press-releases/2015/ 04/10/strata-partners-with-yale-new-haven-health-system-to-reduce-cost-by-im proving-quality. Accessed on: May 23, 2016.

[14] R. Palmerola *et al.*, "Surgical complications and their repercussions," *J. Endourology*, vol. 30, no. S1, pp. S–2, 2016.

[15] A. Güldner, P. M. Spieth, and M. G. de Abreu, "Non-ventilatory approaches to prevent postoperative pulmonary complications," *Best Practice Res. Clinical Anaesthesiology*, vol. 29, no. 3, pp. 397–410, 2015.

[16] G. Ottino *et al.*, "Major sternal wound infection after open-heart surgery: A multivariate analysis of risk factors in 2,579 consecutive operative procedures," *Ann. Thoracic Surgery*, vol. 44, no. 2, pp. 173–179, 1987.

[17] L. Gallart and J. Canet, "Post-operative pulmonary complications: Understanding definitions and risk assessment," *Best Practice Res. Clinical Anaesthesiology*, vol. 29, no. 3, pp. 315–330, 2015.

[18] G. Luc, M. Durand, L. Chiche, and D. Collet, "Major post-operative complications predict long-term survival after esophagectomy in patients with adenocarcinoma of the esophagus," *World J. Surgery*, vol. 39, no. 1, pp. 216–222, 2015.

[19] S. N. Hemmes, A. S. Neto, and M. J. Schultz, "Intraoperative ventilatory strategies to prevent postoperative pulmonary complications: A meta-analysis," *Current Opinion Anesthesiology*, vol. 26, no. 2, pp. 126–133, 2013.

[20] R. G. Johnson, A. M. Arozullah, L. Neumayer, W. G. Henderson, P. Hosokawa, and S. F. Khuri, "Multivariable predictors of postoperative respiratory failure after general and vascular surgery: Results from the patient safety in surgery study," *J. Amer. College Surgeons*, vol. 204, no. 6, pp. 1188–1198, 2007.

[21] R. H. Mehta *et al.*, for the Society of Thoracic Surgeons National Cardiac Surgery Database Investigators, "Bedside tool for predicting the risk of postoperative dialysis in patients undergoing cardiac surgery," *Circulation*, vol. 114, no. 21, pp. 2208–2216, 2006.

[22] I. K. Toumpoulis, C. E. Anagnostopoulos, D. G. Swistel, and J. J. DeRose, "Does EuroSCORE predict length of stay and specific postoperative complications after cardiac surgery?" *Eur. J. Cardio-Thoracic Surgery*, vol. 27, no. 1, pp. 128–133, 2005.

[23] H. F. Elkhenini *et al.*, "Using an electronic medical record (EMR) to conduct clinical trials: Salford lung study feasibility," *BMC Med. Informatics Decision Making*, vol. 15, no. 1, p. 8, Feb. 2015.

[24] R. Amarasingham *et al.*, "Consensus statement on electronic health predictive analytics: A guiding framework to address challenges," *eGEMs*, vol. 4, no. 1, p. 1163, 2016.
[25] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Sci. Rep.*, vol. 6, 2016, Art. no. 26094.
[26] P. Schulam, F. Wigley, and S. Saria, "Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery." in *Proc. Assoc. Advancement Artif. Intell.*, 2015, pp. 2956–2964.
[27] S. Saria and A. Goldenberg, "Subtyping: What it is and its role in precision medicine," *Intell. Syst., IEEE*, vol. 30, no. 4, pp. 70–75, Jul./Aug. 2015.
[28] J. Wiens, W. N. Campbell, E. S. Franklin, J. V. Guttag, and E. Horvitz, "Learning data-driven patient risk stratification models for clostridium difficile," in *Open forum infectious diseases*, vol. 1, no. 2. London UK: Oxford Univ. Press, 2014, Paper ofu045.
[29] M. J. Rothman, S. I. Rothman, and J. Beals, "Development and validation of a continuous measure of patient condition using the electronic medical record," *J. Biomed. Informat.*, vol. 46, no. 5, pp. 837–848, 2013.
[30] G. D. Finlay, M. J. Rothman, and R. A. Smith, "Measuring the modified early warning score and the Rothman index: Advantages of utilizing the electronic medical record in an early warning system," *J. Hospital Med.*, vol. 9, no. 2, pp. 116–119, 2014.
[31] E. Bradley, O. Yakusheva, L. I. Horwitz, H. Sipsma, and J. Fletcher, "Identifying patients at increased risk for unplanned readmission," *Med. Care*, vol. 51, no. 9, p. 761, 2013.
[32] G. L. Piper, L. J. Kaplan, A. A. Maung, F. Y. Lui, K. Barre, and K. A. Davis, "Using the Rothman index to predict early unplanned surgical intensive care unit readmissions," *J. Trauma Acute Care Surgery*, vol. 77, no. 1, pp. 78–82, 2014.
[33] "Epic electronic medical record," [Online]. Available: http://www.epic.com/, Accessed on: May 1, 2016.
[34] "Strata qvi," [Online]. Available: http://www.stratadecision.com/qvi, Accessed on: Aug 1, 2016.
[35] R. G. Brindis, S. Fitzgerald, H. V. Anderson, R. E. Shaw, W. S. Weintraub, and J. F. Williams, "The american college of cardiology-national cardiovascular data registry (ACC-NCDR): Building a national clinical data repository," *J. Amer. College Cardiology*, vol. 37, no. 8, pp. 2240–2245, 2001.
[36] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statistical Softw.*, vol. 33, no. 1, pp. 1–22, 2010. [Online]. Available: http://www.jstatsoft.org/v33/i01/
[37] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: http://CRAN.R-project.org/doc/Rnews/
[38] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, Aug. 13, 2016, pp. 785–794.
[39] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

Authors' photographs and biographies not available at the time of publication.