International Journal of Clinical & Medical Informatics

Research Article | Vol 1 Iss 1

Identifying Hospital Patient Safety Problems in Real-Time with Electronic Medical Record Data Using an Ensemble Machine Learning Model

Michael Li[*], Drew Ladner, Susanne Miller and David Classen

Pascal Metrics Inc., USA

[*]**Corresponding author:** Michael Li, Vice President, Department of Applied Science, Pascal Metrics Inc., 1025 Thomas Jefferson Street NW, Suite 420 East, Washington DC, USA, Tel: +1 202 684 2036; E-mail: michael.li@pascalmetrics.com

Abstract

Objective: There is a need for an electronic measurement and detection tool for patient safety problems in current US hospital systems. We hypothesized that an automated machine learning model using patient clinical data during inpatient stays could stratify patient safety adverse event risk and predict safety problems at the individual patient level with superior accuracy. In this study, we developed a patient safety harm prediction algorithm using machine learning and electronic medical record (EMR) patient data that can be implemented for real-time use to support patient safety and quality improvement initiatives.

Methods: Historical EMR clinical data from hospitals within the Pascal Metrics Patient Safety Organization (PSO) was used. The EMR data included a total of 10,315 inpatient discharges over an eight-month period from October 2012 to June 2013. The outcomes were clinician-authenticated adverse events collected in the Pascal Metrics Risk Trigger® Monitoring (RTM) PSO system. The predictive algorithm was developed using an ensemble machine learning approach that combines bagging, boosting and random feature methods.

Results: The machine learning model had a C-statistic of 0.91 in the training set and a C-statistic of 0.88 in the validation sample. The adverse event risk score at 0.1 levels can identify 57.2% of adverse events (sensitivity) with 26.3% accuracy (PPV) from 9.2% of the validation sample. The adverse event risk score of 0.04 can identify 85.5% of adverse events with 11.5% PPV from 31.4% of the validation sample.

Conclusion: We developed an ensemble machine learning algorithm using inpatient EMR data and clinically validated adverse event outcomes to automatically predict all-cause harms at the individual patient level with a superior accuracy.

Keywords: ***Adverse event risk prediction; All-cause harm measurement; EMR data; Ensemble machine learning; Patient safety and quality improvement***

Citation: Michael Li, Identifying Hospital Patient Safety Problems in Real Time with Electronic Medical Record Data Using an Ensemble Machine Learning Model. Int J Clin Med Info 2018: 1(1) 43-58.

Received Date: May 09, 2018; **Accepted Date:** July 10, 2018; **Published Date:** July 17, 2018

Introduction

Widespread preventable injury to hospitalized patients persists more than 18 years after the Institute of Medicine (IOM) report, "To Err is Human," brought it to international attention. In fact, medical error was recently cited as the third-leading cause of death in US hospitalized patients [1,2]. A recent study suggested that patient safety problems may lead to the death of more than 400,000 inpatients in the Unites States every year and injure another 8 million inpatients [3]. This magnitude of harm will likely persist until it is measured effectively, efficiently, and consistently by every hospital. Among the barriers to measurement of harm are the lack of widely embraced standard definitions for adverse events, the perceived costs of data collection, and the use of largely manual methods to identify adverse events in the United States through the reliance on primarily voluntary event reporting. All of these barriers can be lowered with better methods in order to measure harm more effectively, efficiently, and consistently through the use of EMR data.

EMRs have the potential to improve patient safety, but their implementation and use has met with unintended consequences and new safety concerns [4-6]. A recent survey by the US Office of the National Coordinator found that more than 96% of US hospitals have implemented an EMR system [7]. However, most hospitals do not use the EMR to directly measure patient harm. Instead, measurement is more often based on electronic voluntary event reporting systems and automated coding of medical records at discharge. One clinical area, infection prevention surveillance, has harnessed the power of EMR data to improve detection of safety problems. Electronic surveillance of infections has enabled practitioners to monitor all hospitalized patients efficiently, effectively, and using standard definitions. Automated identification of laboratory diagnostics, such as positive blood culture results, enabled improved detection. This approach greatly enhanced management and reduction of serious infections, such as central line associated blood stream infections. Further improved measurement of infections has accelerated efforts to reduce such harm. But some hospitals are moving beyond infection control to measure safety more broadly with EMR-enabled all-cause harm surveillance, presaging the beginning of a new trend to begin using EMRs to measure, monitor, and manage patient safety more broadly and in real-time [8].

Currently, most hospitals use voluntary event reporting and administrative coding of medical records to measure patient safety [9,10]. This approach reflects both operational convenience and regulatory requirements. Most hospitals have purchased voluntary event reporting systems to manage adverse events, and all hospitals are required to report to the Centers for Medicare and Medicaid Services (CMS) using the hospital-acquired condition (HAC), which is measured using administrative billing codes. Incidents, complaints, and claims reporting systems are less suitable for counting adverse events, because the number of adverse events strongly depends on the willingness of healthcare providers and patients to report them. Thus only 3-5% of the adverse events detected from inpatient records are reported by healthcare providers in hospitals. In addition, the denominator, the related number of patients, is difficult to determine. These systems are therefore inadequate to count the actual number of incidents [11-13]. In addition, the CMS approach of using administrative data to measure safety has been shown to miss more than 90% of safety problems [14].

From a scientific point of view, the clinical review of patient records is the most thoroughly studied method used to measure the prevalence of adverse events and is considered to yield clinically valid documentation. In addition, it has the highest sensitivity and specificity of all methods [10-12]. However, it historically has been a completely manual process, is the most labor-intensive and expensive, and consequently has not been widely relied upon in hospital operations. The perception of its prohibitive cost and assumed lack of scalability has resulted in questioning whether clinical review can play a role in machine learning and advanced AI methods for detecting patient safety problems. Overall, there is a need for a comprehensive measurement and detection method for adverse events in current US hospital systems that is more effective than prevailing methods in identifying harm - namely, event reporting and administrative coding - but, through the automated use of EMR data, which is more efficient than traditional clinical review.

This movement to real-time measurement of patient safety using EMR data allows for the development of large data sets that can be used to create predictive analytic methods for patient safety that can predict patient safety problems before they occur. Many types of clinical predictive methods are already in use for anticipating outcomes such as clinical deterioration, hospital readmission, heart failure deterioration, surgical complications, and the list continues to grow rapidly. The aforementioned definitional, cost, and validation challenges associated with patient safety measurement have impeded similarly rapid advances in predictive analytics. This paper outlines the approach used to develop a predictive algorithm for patient safety problems in hospital patients that addresses these challenges using rich EMR data, automation-enabled and scalable clinical review, and clinically validated safety outcomes data at the adverse event category level.

In order to build a better predictor, this study used a machine learning approach. Traditional statistical modeling focuses more on inference and requires model fit, calibration, and model assumption diagnosis. Machine learning focuses only on model prediction by manipulating each of the three components in the following relation of Y=f(X), where Y is the response or target variable (s), X the predictor or feature set and f the function or systematic mapping from X to Y. Model fit, estimation and diagnosis are not important for machine learning models given that prediction accuracy is the main goal and criteria. Both statistical models and machine learning models can provide predictions, but the strength of machine learning in predicting clinical outcomes is in its ability to handle a large number of input predictor variables, an automated way to create and select features, the ability to learn or model complex relationships between X and Y, and, most importantly, better prediction accuracy as a result. Intuitively, machine learning, such as ensemble machine learning, and the use of artificial neural networks can capture more complex structures between X and Y and even be able to strengthen weaker signal detections. In comparison, typical statistical models such as multivariate logistic and survival models can only provide simpler or log-linear relationships.

The important aspect of machine learning is to create features or variables that can be used for model training. Good features can extract information and signals distinctively and specifically as different risk factors. Poor feature constructs are more aggregated and less expressive for underlying risk factor information. There are different schools of thought regarding how the features should be generated such as using an automated machine method or using an interactive exploration process with human judgment. Given the nature of clinical data, the critically important role of clinical validation to support sound epidemiology, and the small sample sizes of data often found at the adverse event category level, the co-authors view the second approach as more appropriate. It also allows us to leverage commonly used normal reference ranges for many clinical measures and some Institute for Healthcare Improvement Global Trigger Tool (GTT) trigger threshold levels from the literature. For the ensemble model construct, we used a semi-automated approach to select different classifiers, which can add more model diversity, improve accuracy, and avoid overfitting.

We hypothesized that an automated machine learning model using patient clinical data during inpatient stay could stratify adverse event risk and predict harm at the individual patient level with excellent accuracy. In this study, we developed an all-cause harm prediction algorithm using machine learning and EMR patient data that can be implemented in real-time use.

Materials and Methods

Study Population

Hospital inpatient EMR data contained within the Pascal Metrics PSO was used, with a total of 10,315 inpatient discharges over an eight-month period from October 2012 to June 2013. As provided by US federal statue and regulation, Pascal Metrics PSO member data can be analyzed and utilized by Pascal for purposes of patient safety and quality improvement. All patients who had clinical data (such as lab and vital sign results) were included in the study population.

Adverse event outcome collection

Data for the model was obtained from physician-authenticated adverse events documented in the Pascal Metrics Risk Trigger® Monitoring (RTM) system; a cloud-based, multi-tenant real-time patient safety surveillance system that exists within the Pascal Metrics PSO. The system received clinical data via Health Level 7 (HL7) based interfaces from the hospital EMRs. The patient data was normalized, mapped to standards, and evaluated for matches to trigger definitions within the system, which are based on the triggers identified in the GTT [15]. Positive triggers were then presented to users in the application. Clinical quality analysts reviewed the list of triggers on a daily basis to determine if an adverse event occurred for the patient. The team used a set of standardized guidelines detailing how to systematically review a specific trigger and then to classify the event if an adverse event was discovered. The quality analyst then documented the adverse event, classified the type of event (for example adverse event related to medication, patient care, perinatal care, surgical care or healthcare associated infection), and assigned a severity score based on the NCC-MERP scale (which is a 9-point scale with values from A-I). Completed documents were then authenticated by a physician reviewer to confirm that the events as documented are clinically valid and reflect adverse events related to the medical care of the patient and not as a result of the patient's disease.

Outcome data description

The primary outcomes used were adverse events that occurred during the inpatient stay in the hospital. Only authenticated adverse events were included in the modeling. Any adverse events that were present on admission were excluded from the analysis. The outcomes exclude events with a severity level of A through D since by definition these events did not reach the patient. However, the events with a severity level of E through I were included since these scores represent events that reached the patient. Of the total 431 adverse events in the study data set, 37.6% of them were at the E level of severity, 52.9% at the F level of severity, and the remaining 9.5% were at a severity level of G and above.

All of these inpatient adverse events were represented by forty-eight categories of adverse events. The most frequent category was "Medication-related glycemic events" (11.4%), followed by "Medication-related bleeding" (10%) 002C "Respiratory complications related to surgery or procedure" (8.4%), "DVT/PE" (6.3%), "Respiratory infection (non-ventilator associated)" (5.8%), "Medication-related delirium, confusion or over-sedation" (5.1%), and "Post-operative/Post-procedure wound infection" (4.2%). Given the relatively small volume of the outcomes in each category, we did not predict specific types of adverse events but, rather, all adverse events.

Data processing and feature engineering

In order to predict inpatient adverse event outcomes, we used patient clinical data collected during patient stays. Patient diagnoses and procedure codes were not used since these billing codes are applied later after patient discharge and clinical data may represent patient conditions more accurately. Inpatient EMR data were used to create many features that can used as machine learning model inputs. Data cleaning and data exploration were performed before our manual feature engineering. Feature engineering is the process of creating new input features or predictor variables for machine learning models. We sought to create features that can extract information or isolate risk factors for adverse events by using existing medical knowledge and data distribution properties. For many clinical measures, there exist normal ranges (and with variations) in medical literature. We sought to use those normal ranges to guide our feature coding. We also used the clinical measure distributions to guide feature definitions. Given that the data volume for this study was limited, we had to ensure each indicator had sufficient volume to determine statistical significance.

The data elements that were used for coding features included lab results, vital sign measures, medication usage, hospital utilization and patient movement within hospital as shown in (Table 1). There were 58 clinical measures from lab tests, vital signs, microbiology, nursing, surgery, radiology, medications and others.

Lab Test [16]	Albumin, Bilirubin, Blood Urea Nitrogen (BUN), Creatinine, Creatinine Phosphokinase (CPK), Arterial Carbon Dioxide partial pressure (pCO2), Glucose, Creatinine Kinase-MB (CKMB), CKMB Index, Troponin, International Normalized Ratio (INR), N-terminal pro B-type natriuretic peptide (NT-proBNP), Prothrombin Time (PT), Arterial pH, Chloride, Potassium, Sodium, Aspartate Aminotransferase (AST), Lactic Acid Dehydrogenase (LDH), Hematocrit, Hemoglobin, Platelet, Partial Thromboplastin Time (PTT), White Blood Count (WBC)
Vital Sign [6]	Temperature, Heart rate, Respiratory rate, Diastolic blood pressure, Systolic blood pressure, Oxygen saturation
Another Event or Test [10]	Surgery event, Surgery time, Intubation date, Medical restraint initiation date, C-difficile test result, radiology tests order (CT, ultrasound, VQ scan) for Emboli or DVT, Fecal Occult Blood test result, Braden total score, Fall Risk score, Fraction of inspired oxygen (FiO2)
Medication [10]	Antiemetic, Anticoagulant, Oxytocin, Terbutaline, Vitamin K, Flumazenil, Diphenhydramine, Narcan, Vasopressors, total medications
Check-in Event [5]	Location change, room and bed change, physician change, ICU change, unplanned ICU
Hospital Use [2]	Prior hospitalization and LOS

Table 1: EMR Data Elements Used for Model Development.

All data values were converted into numerical values, including the contents of text values. Text values such as “<0.1” or “>600” were converted to their truncated numbers. Distribution of every data element was plotted and examined. As an example, (Figure 1) shows the lab value distribution for Albumin, BUN, Creatinine and Glucose.

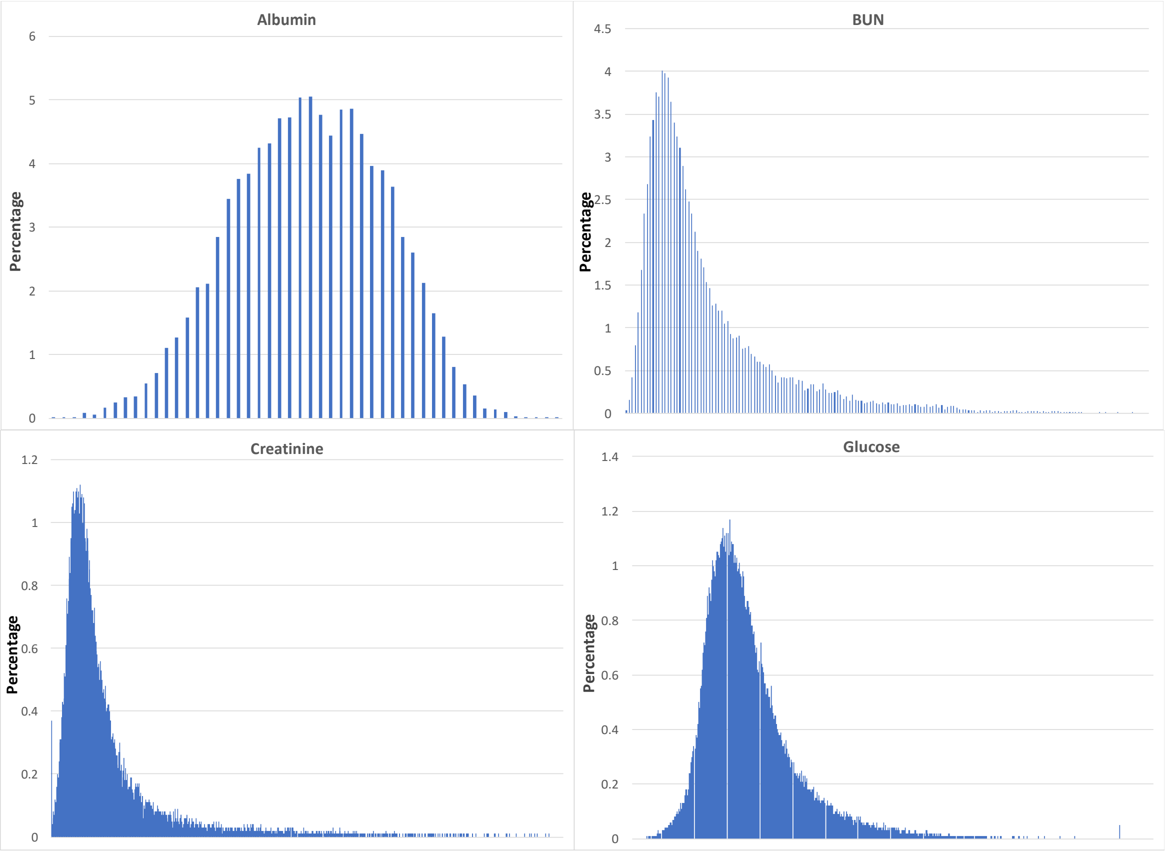

Figure 1: Lab Value Distribution: Albumin, BUN, Creatinine and Glucose

For clinical data elements such as lab and vital sign results, we first created maximum, minimum, first, and last readings and a number of result readings for each patient. Then we created level category indicators based on the distribution or typical normal reference ranges. These indicators can measure both tail ends of clinical event distribution such as blood pressure, or glucose level, which measure different types of condition risks. For a clinical event with multiple readings during a patient visit, we created a rate of change over time and then use the distribution to create change indicators.

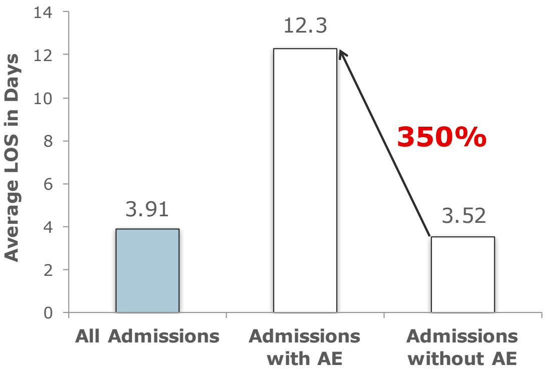

Figure 2: Average LOS of AE vs. Non-AE patient visits.

This method can generate from four to over 20 discrete indicators for a lab measure, which provided a total of 467 features or variables. These features capture different levels and variations of each clinical data element for a patient. One of the predictors is patient length of stay, which had a strong correlation with adverse events. The average length of stay for patients with adverse events is 12.3 days versus 3.5 days for patients without adverse events, as shown in (Figure 2).

We used this manual and labor-intensive approach supported by clinical judgment to create features, instead of pure machine-generated features. This method incorporates certain existing medical knowledge such as normal condition values, which are easier to measure and interpret as risk factors once the model is developed. In fact, this method is quite consistent with the GTT methodology [15]. Indeed, as a starting point, we sought to use the same threshold values for clinical events as triggered by the GTT, wherever applicable, based on clinical and scientific judgment.

Ensemble machine learning method

Ensemble models can improve the prediction performance of a single classifier [17-28]. The term ‘‘ensemble” refers to the combination of multiple classifiers. Most common ensemble machine learning models use decision trees as base learners. We designed and implemented a unique ensemble method which uses multivariate logistic model as the base learner computing engine and three ensemble meta-algorithms on top of the base learner. The ensemble model predicts a target value as an average or a vote of the predictions from several individual classifiers or models. Contrasted with the traditional single multivariate logistic model prediction that is commonly used for outcome prediction in clinical research, an average of several predictions is often more accurate and stable than a single model prediction, even though each individual model may be less accurate or weaker. We sought to have several models that can predict independently of the others by using different ensemble meta-algorithms. We trained a total of fifteen classifiers and gave them equal weighting for the final prediction of the targeted outcome adverse event.

The fifteen classifiers were trained using three types of machine learning ensemble meta-algorithms: bootstrapping aggregation or bagging, boosting, and the random subspace method [18-28]. We used these 15 models with three ensemble approaches to create independent and heterogeneous classifiers to improve prediction accuracy, i.e., capturing small outcome sub-categories, thereby reducing bias and variance of the predictions. The total number of individual classifiers can be increased and other ensemble methods can be used as well. For simplicity, we decided to have only fifteen classifiers which statistically provide a good estimate of the mean for the individual voting scores.

We had two classifiers using the bagging method [16,18,23,28], where each training set was randomly sampled with replacement. Then each subset is trained by the multivariate logit model with the same feature set to obtain three bagging classifiers. The logistic model used backward variable selection with p-value of 0.05 for features or variables to be included in the final model.

There are six classifiers using a two-step boosting method [20,21,23,27,28]. Boosting is an ensemble technique that trains the models sequentially. In boosting, classifiers are constructed on weighted versions of the training set, which are dependent on previous training results. The misclassified cases from the first classifier or model were over weighted and then used as the targets for training in the second model using the same feature set but reselected by machine. The second classifier is boosted on the reweighted training set. The second classifier is better in predicting the adverse event cases that were missed out from the first one, which may not predict well for most common targets, but is more adapted to predict the abnormal or rare cases in second classifier. The misclassified cases from the second classifier were over weighted again and then used as the targets for training in the third model using the same feature set but reselected by machine. The third classifier is the second boost on the reweighted training set.

In the random subspace method, classifiers are constructed in random subsets of the data feature space [22,16,28]. Different models can be trained using different feature sets to provide different diverse classifiers. We have seven classifiers using random subspace method. There are many correlated features in our data as multiple clinical conditions can be related to a medication error or an infection. Therefore, one may obtain better classifiers in random subspaces than in the original feature space and provide more independent classifiers for a final superior risk score.

Data Characteristic	Entire Sample (N = 10316)	Training Set (N = 6876)	Validation Set (N = 3440)	P-Value
Adverse Event	4.19%	4.10%	4.36%	0.5354**
Mortality	1.58%	1.64%	1.45%	0.4659**
LOS	3.784	3.77	3.812	0.7642*
Female	56.30%	56.50%	55.90%	0.5914**
Race - Black	8.30%	8.20%	8.60%	0.4853**
Race - White	71.80%	72.50%	70.60%	0.0397**
Albumin	1.988	2.003	1.959	0.2443*
Bilirubin	0.372	0.376	0.364	0.6107*
BUN	17.7	17.9	17.4	0.2195*
Creatinine	0.938	0.95	0.915	0.1731*
Glucose	108	108.3	107.4	0.9650*
Hematocrit	31.9	32.1	31.7	0.4030*
Hemoglobin	10.6	10.6	10.5	0.0900*
INR	0.781	0.79	0.765	0.1493*
Platelet	205.6	206.7	203.5	0.1761*
Potassium	3.219	3.247	3.162	0.0141*
Troponin	0.044	0.046	0.04	0.4689*
WBC	9.08	9.11	9.03	0.5800*
Temperature	95.4	95.5	95.4	0.8735*
Respiratory Rate	17.8	17.8	17.8	0.3282*
Pulse	80	80	80	0.9975*
BP Diastolic	68.6	68.5	68.7	0.3905*
BP Systolic	125	124.8	125.4	0.1994*
Oxygen	93.9	93.9	93.9	0.9580*
Braden Score	18.5	18.5	18.6	0.2687*
Intubation	4.80%	4.90%	4.70%	0.7408**
# of Medications	16.7	16.8	16.5	0.2412*

Table 2: Profile for Selected Data Elements.

*P-value for two sample Chi-square test

**P-value for two sample t-test

Combining all classifiers from bagging, boosting and random subspace method, we obtain our final adverse event prediction score by averaging all fifteen individual predictions in the training data. Comparing to traditional statistical modeling approach using a single multivariate logistic model to predict, our ensemble machine learning model uses many multivariate logistic models, or "the wisdom of crowd," to predict adverse events.

All data processing, analysis, feature engineering, ensemble model coding, testing and validation were performed using the SAS® software version 9.2.

Results

Description of the derivation and validation cohorts

The total data sample was randomly sampled into two cohorts, two-thirds (6,876) for model training and one-third (3,440) for model validation. Selected patient clinical and demographic characteristics for the entire sample, as well as training and validation cohorts are shown in (Table 2). A two-sample t-test was used for all continuous measure comparison, and two sample Chi-square tests were used for categorical measure comparison. The average length of stay was 3.78 days; 56.3% patients were female; and 71.8% patients were white. There were no statistically significant differences in all continuous measures between the test and validation data set based on lab test and vital sign values. There were no significant differences in adverse event, mortality, gender and race measures, with the exception of the white population in the training set being marginally higher than in the validation set.

Predictors for adverse events

There is a total of 467 variables in the final feature set that we used to select variables for model training. (Figure 3) shows a bi-plot of positive predictive value (PPV) and sensitivity values with respect to adverse event outcome for all 467 features. The PPV measure here shows the proportion of positive feature values that actually have adverse events. Sensitivity measures the proportion of a feature being positive to identify correctly those patients who have adverse events. Typically, there is a trade-off between PPV and sensitivity for many features. There were 112 features that have a relatively high PPV of more than 10% and 172 features with a sensitivity value more than 10%.

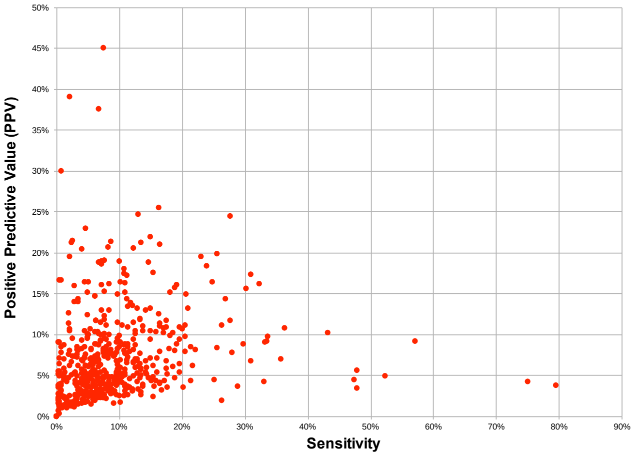

Figure 3: Sensitivity and PPV Bi-plot for All Features Used in Training.

To measure and compare how features correlated with the outcome measure, we also calculated accuracy, an F1 Score, and a Matthews Correlation Coefficient (MCC) for each feature [29]. Accuracy is the percentage of total correct predictions (both true positive and true negative) from total predictions. The F1 score is a composite measure between a predictor and outcome measure, which is the harmonic mean of sensitivity and PPV. The MCC is similar to a correlation coefficient between a predictor and target outcome with a value between −1 and +1. (Figure 4) shows there are a very high correlation between F1 score and MCC. There are small differences in low F1 score and low MCC score (close to zero).

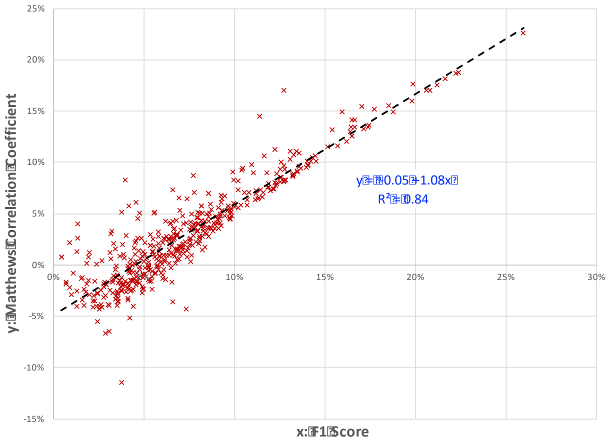

Figure 4: F1-Score and MCC Bi-plot for All Features.

The mean, standard deviation, maximum, and minimum of the PPV, NPV, sensitivity, specificity, accuracy, F1 Score and MCC of all 467 features are provided in (Table 3). There were many predictors or features that have small correlations or predicting power with respect to adverse event outcome. Ensemble machine learning can use all these small predictors to form one much stronger predictor by capturing a more complex relationship structure and a group of predictors simultaneously.

Measure	Sensitivity	PPV	Specificity	NPV	Accuracy	F1-score	Matthews Correlation Coefficient
Mean	10.40%	7.70%	92.60%	95.90%	89.10%	7.40%	3.00%
Standard Deviation	9.10%	5.70%	8.00%	0.30%	7.40%	4.20%	5.00%
Maximum	79.40%	45.10%	99.90%	97.60%	95.80%	26.00%	22.60%
Minimum	0.00%	0.00%	13.30%	93.40%	16.10%	0.50%	-11.50%

Table 3: Feature Profile Statistics with Respect to Adverse Event Outcome.

Model performance

Fifteen multivariate logit classifiers were trained using bagging, boosting and random subspace method on 6,876 testing sample. The backward elimination method was used to automatically select variables in each classifier. Different ridging options (Absolute and Relative) were used to provide improvement and variations in variable selection and prediction [30]. There were 254 variables in the final trained ensemble model from the initial 467 feature set. The area under the receiver operating characteristic (ROC) curve or C-statistic of the ensemble model in the training set was 0.9058 and 0.8806 in validation set as shown in (Figure 5 and Figure 6). The C-statistic shows the trade-off between sensitivity (true positives) and 1-specificity (false negatives) at all possible cutoff values of the prediction risk score. A value of 0.5 C-statistic indicates a model is no better than chance at making a prediction and a value of 1.0 indicates a model has perfect prediction. Models are typically considered reasonable when the C-statistic is higher than 0.7 and strong when C-statistic exceeds 0.8. Currently there is no adverse event prediction algorithm using EMR data with clinical validation in the literature with which we can compare our study. Published EMR-based readmission models provide perhaps the closest, albeit imperfect, comparison. The C-statistics for those models were between 0.68 and 0.83 [31-33].

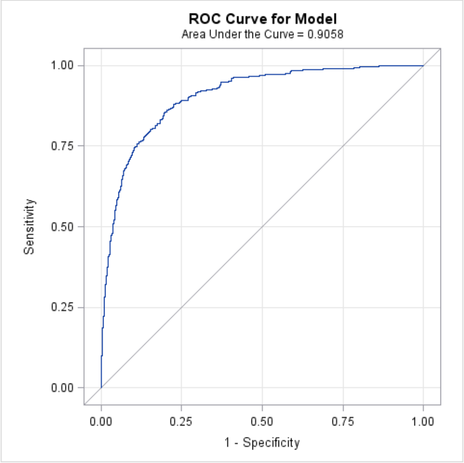

Figure 5: Training Set Performance.

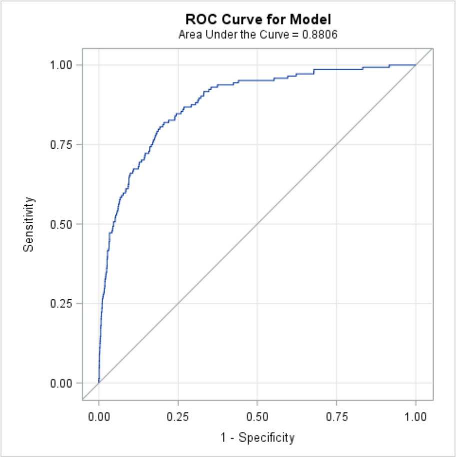

Figure 6: Validation Set Performance.

The model prediction score had a much higher PPV and sensitivity with regard to adverse event outcomes relative to individual features, which shows the predictive power of the machine learning model. (Table 4) shows the PPV and sensitivity of the machine learning model score at selected threshold levels for the validation sample. The adverse event risk score at the 0.1 level can identify 57.2% of adverse events (sensitivity) with 26.3% accuracy (PPV) from 9.2% of the validation sample patients. The adverse event risk score of 0.04 can identify 85.5% of adverse events with a 11.5% PPV from 31.4% of the validation sample patients. The PPV is 36.6% when the risk score is 0.15, and 40.7% when the score is 0.2. The model can identify most of adverse event cases in the validation sample with a good “hit rate” or PPV. Assuming the use of 4% as our decision rule to trigger an adverse event alert, the 31.4% validation sample patients are triggered positive, so providers can spend time on only 31.4% of patients for exploring intervention and adverse event review and documentation versus 100% or, that is, all patients. This is the resource saving benefit of our machine learning algorithm. Using this 4% rule, most, or 85.5%, of adverse events can be identified. At this level of sensitivity, the PPV is 11.5% PPV.

Risk Score	% of Patients	PPV	Sensitivity
0.4	0.20%	62.50%	3.40%
0.3	1.00%	57.10%	13.80%
0.2	3.30%	40.70%	31.70%
0.15	5.00%	36.60%	43.40%
0.1	9.20%	26.30%	57.20%
0.08	13.20%	20.90%	65.50%
0.06	19.20%	16.20%	73.80%
0.05	24.00%	14.20%	80.70%
0.04	31.40%	11.50%	85.50%
0.03	43.40%	9.00%	92.40%
0.02	73.00%	5.70%	98.60%

Table 4: Model Performance in the Validation Sample.

The risk score distribution is similar to a Pareto distribution, and the Pareto principle or "80-20 rule" applies here: 80.7% of adverse events affected 24% of patients. The predictive model is valuable not only in predicting individual patient risk but also in identifying and segmenting all patients in high, medium, and low risk categories for care and outcome documentation and measurement.

Discussion and Conclusion

The patient safety predictive model presented here can be operationalized in several ways to help hospitals reduce patient harm. First, it can be used in real-time to identify individual patients who are most likely to have adverse events throughout their entire hospital stay. The risk score from the model can be used as a stratification tool to segment all patients into high, medium and low risk categories enabling hospital staff to allocate resources and prioritize adverse event prevention activities to improve patient safety and reliability. Second, the model can be implemented and used dynamically within a software application so that new pieces of patient information can be updated in real-time and reflect adverse event risk changes over time [34,35]. Third, and importantly, this adverse event risk algorithm can be a valuable tool to support hospital-wide adverse event measurement and documentation as part of holistic patient safety and quality improvement programs to support the safety and reliability of care.

The GTT that generates validated adverse event outcomes data is widely used for measuring adverse events in the hospital setting [9,36] and has been demonstrated to be superior in identifying harm as compared to other methods [9]. However, as typically implemented, the GTT only randomly samples 20 patients per hospital per month for chart review because of the prohibitive cost of conducting clinical review of all patients. Further, it is highly retrospective (i.e. often by weeks or months) and therefore lacks actionability.

The machine learning adverse event algorithm we developed offers a solution to resolve the tension between two schools of thought: (i) those who, on the basis of sound epidemiological principles, hold that measurement of outcomes must be clinically validated by physician-authenticated review and (ii) those who, on the basis of the high financial cost and potential unreliability introduced by the human factor, seek to minimize if not eliminate altogether clinical review by applying technology, perhaps even incorporating a highly advanced machine learning and AI methods as has this study.

This method addresses the concerns of both schools of thought. First, sound epidemiological principles are applied both in documenting the authenticated outcome data upon which the algorithm is constructed as well as in the clinical discovery and decision-making that occurs when the stratified patients are identified for potential intervention; doing so satisfies the school of thought rejecting the elimination of clinical judgment. Second, the software algorithm screens every patient in real-time with robust risk assessment; doing so satisfies the school of thought rejecting the inefficiency and ineffectiveness of labor-intensively conducting clinical review on every patient constantly. Indeed, the tool enables clinicians to identify potential risk with excellent prediction in every patient at low cost continuously around the clock, demonstrating scalability across the largest health provider environments. In sum, this novel method brings together the foundation of clinically validated safety outcomes and the promise of machine learning and advanced AI techniques.

Furthermore, for those still concerned about the absolute cost of clinical review, this method can generate financial benefits exceed those costs. Because the tool is firing against streaming data in real-time, successful interventions (e.g. reducing length of stay, readmissions, and downstream utilization across the continuum) can be executed while the patient is receiving care and generate contribution to financial margin. Based on the literature calculating the cost of harm, the benefits generated can exceed the cost of the limited clinical review, thereby generating both clinical and financial value [8]. This value is amplified for those providers and health systems bearing payment risk for financial outcomes. Finally, combining the demonstrated efficacy of detective clinical triggers and a real-time predictive algorithm based on EMR data can provide a more comprehensive approach for hospitals seeking to reduce patient harm and related cost: richer adverse event documentation; more accurate and timely measurement; and clinically relevant and useful monitoring and management that all together can become foundational for the next generation of patient safety, quality improvement, and risk management.

To our knowledge, no previous study demonstrates the prediction of all-cause harm in the hospital setting based on clinically validated EMR-based adverse event outcomes. Possibly due to the difficulty in accessing adverse event outcome data, most studies applying prediction in patient safety have limited scope to the prediction of conditions. The Rothman Index [34] can predict patient deterioration dynamically with a single score. However, their model used patient one-year post discharge mortality as the targeted outcome to develop a heuristic model with 26 clinical measurements. Bihorac et al. [35] used a generalized additive model (GAM) to predict risk probabilities for eight major postoperative complications for inpatient surgeries. Our model used all-cause harm outcomes and 51 clinical measurements in addition to patient demographic and admission activity measures to train our model. It is possible that some critically ill patients could be caused by adverse events, but not all patients with adverse events are critically ill or necessarily deteriorate. Finally, there are an increasing number of studies that seek to predict adverse events but are significantly limited by (i) the lack of EMR-based adverse event outcomes data, for which imperfect proxies such as readmission, mortality, and morbidity are substituted; (ii) the lack of application of machine learning and advanced AI techniques; or (iii) the lack of clinical validation which, even if it can be proven to be possible, will find limited adoption in the near future vis a vis a method that includes clinical review cost-effectively.

There are limitations to this study. First, the data for this study was based on data within a PSO that represented a limited number of community hospitals. The model performance for other community or academic hospitals may not be as good as the current facilities. Second, the model performance can be improved by using a larger sample of patient data. By using higher data volume in a subsequent study, more clinical measures can be used due to having sufficient volume, subject to evaluation and then utilized for adverse event prediction. Also, some specific types of adverse events had quite small volumes, such as "Clostridium difficile medication associated infection", "Medication-related coagulopathy" or "Contrast dye-related acute renal injury". Prediction for such specific type adverse events can be improved as these outcome data volumes accumulate. Third, our current construct of ensemble learning was a priori in the sense that each base learner was constructed to be independent and diversified without a loss function associated with the whole ensemble for base learner improvement, which is a similar approach as that used by the Netflix Challenge winner [37]. Decision tree-based models such as the gradient boosting machines (GBMs) [38,39] and XGBoost [40,41] can also be applied. These methods have shown considerable success in many data mining and machine learning challenges recently. However, bigger data and more sophisticated feature engineering are more important, if not equally, for prediction performance than technical machine learning techniques in our opinion.

We present a new all-cause harm adverse event prediction algorithm for hospital inpatient setting that can be used for risk stratification, individual patient adverse prediction, surveillance, and adverse event documentation. This algorithm used an ensemble machine learning model and was trained by using EMR data with clinically validated adverse event outcomes. The model demonstrated an excellent predictive performance in the validation sample. The algorithm can enhance and supplement a trigger based adverse detection method and functions as a patient level prediction tool for adverse events and inpatient adverse event risk screening. Feeding this algorithm with real-time EMR data feeds, this model and extensions thereof-such as to predict specific types of adverse event outcomes-has the capability of being delivered dynamically in real-time to a range of users in a hospital environment and across the continuum of care.

Conflict of Interest

The authors are employees and shareholders of Pascal Metrics Inc., a Patient Safety Organization (PSO) with the mission of improving the safety and reliability of health care worldwide. The Risk Trigger® system is a software service offered as part of the Pascal Metrics PSO solution and is currently deployed at major hospitals and health systems. The modeling methodology described herein to create the adverse event risk score, alternatively named the Global Safety Risk (GSR) Score, was launched in 2012 as an internal research and development initiative to apply machine learning and advanced AI techniques to patient safety, quality improvement, and risk management.

Acknowledgment

The authors express their appreciations to following people who provided engineering and data support for this research, Theodore Nguyen-Cao, Andrew Hicks, and Eric Simmerman.

References

1. Committee on Patient Safety and Health Information Technology; Board on Health Care Services; Institute of Medicine (2012) Health IT and Patient Safety: Building Safer Systems for Better Care. Washington, DC: National

Academies.

2. Daniel M and Makary MA (2016) Medical error-the third leading cause of death in the US. BMJ 353(i2139): 476636183.
3. James JT (2013) A new, evidence-based estimate of patient harms associated with hospital care. Journal of Patient Safety 9(3): 122-128.
4. Koppel R, Metlay JP, Cohen A, et al. (2005) Role of computerized physician order entry systems in facilitating medication errors. Journal of the American Medical Informatics Association 293(10): 1197-1203.
5. Harrington L, Kennerly D and Johnson C (2011) Safety issues related to the electronic medical record (EMR): synthesis of the literature from the last decade, 2000-2009. Journal of Healthcare Management 56(1): 31-44.
6. Magrabi F, Ong MS, Runciman W, et al. (2010) An analysis of computer-related patient safety incidents to inform the development of a classification. Journal of the American Medical Informatics Association 17(6): 663-670.
7. https://dashboard.healthit.gov/quickstats/pages/FIG-Hospital-EHR-Adoption.php
8. Sammer C, Miller S, Jones C, et al. (2017) Developing and evaluating an automated all-cause harm trigger system. The Joint Commission Journal on Quality and Patient Safety 43(4): 155-165.
9. Classen DC, Resar R, Griffin F, et al. (2011) ‘Global trigger tool’shows that adverse events in hospitals may be ten times greater than previously measured. Health Affairs 30(4): 581-589.
10. Murff HJ, Patel VL, Hripcsak G et al. (2003) Detecting adverse events for patient safety research: a review of current methodologies. Journal of Biomedical Informatics 36(1-2): 131-143.
11. Vincent C, Burnett S, Carthey J (2013) The measurement and monitoring of safety. London: The Health Foundation.
12. Thomas EJ and Petersen LA (2003) Measuring errors and adverse events in health care. Journal of General Internal Medicine 18(1): 61-67.
13. Dückers M, Faber M, Cruijsberg J, et al. (2009) Safety and risk management interventions in hospitals. Medical care research and review, 66(6_suppl): 90S-119S.
14. Tsang C, Aylin P, Palmer W (2008) Patient safety indicators: A systematic review of the literature. London, UK: Dr. Foster Unit, Imperial College.
15. Griffin FA and Resar RK (2009) IHI Global Trigger Tool for Measuring Adverse Events, 2nd edn. Cambridge, MA: Institute for Healthcare Improvement.
16. Skurichina M and Duin RP (2002) Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis & Applications 5(2): 121-135.
17. Kittler J, Hatef M, Duin RP, et al. (1998) On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3): 226-239.
18. Breiman L (1996) Bagging predictors. Machine Learning 24(2): 123-140.
19. Breiman L (1996) Arcing classifiers. Annals of Statistics 26.
20. Freund Y and Schapire RE (1996) July. Experiments with a new boosting algorithm. International Conference on Machine Learning 96: 148-156.
21. Schapire RE, Freund Y, Bartlett P, et al. (1998) Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics 26(5): 1651-1686.
22. Ho TK (1998) The Random subspace method for constructing decision forests. IEEE Trans Pattern Analysis and Machine Intelligence 20(8): 832-844.

23. Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning 40(2): 139-157.
24. Dietterich TG (2000) Ensemble methods in machine learning. In International workshop on multiple classifier systems (1-15). Springer, Berlin, Heidelberg.
25. López V, Fernández A, García S, et al. (2013) An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences 250: 113-141.
26. Nasrabadi NM (2007) Pattern recognition and machine learning. Journal of Electronic Imaging 16(4): 049901.
27. Friedman J, Hastie T and Tibshirani R (2001) The elements of statistical learning 1(10). New York, NY, USA: Springer series in statistics.
28. Zhou ZH (2012) Ensemble methods: foundations and algorithms. Chapman and Hall/CRC.
29. Chicco D (2017) Ten quick tips for machine learning in computational biology. BioData Mining 10(1): 35.
30. Mansson K and Shukur G (2011) On ridge parameters in logistic regression. Communications in Statistics-Theory and Methods 40(18): 3366-3381.
31. Amarasingham R, Moore BJ, Tabak YP, et al. (2010) An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Medical care 48 (11): 981-988.
32. Nijhawan AE, Clark C, Kaplan R, et al. (2012) An electronic medical record-based model to predict 30-day risk of readmission and death among HIV-infected inpatients. JAIDS Journal of Acquired Immune Deficiency Syndromes 61(3): 349-358.
33. Kansagara D, Englander H, Salanitro A, et al. (2011) Risk prediction models for hospital readmission: a systematic review. The Journal of the American Medical Association 306(15): 1688-1698.
34. Rothman MJ, Rothman SI, Beals IV J (2013) Development and validation of a continuous measure of patient condition using the Electronic Medical Record. Journal of Biomedical Informatics 46(5): 837-848.
35. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, et al. (2018) MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. Annals of surgery.
36. Hibbert PD, Molloy CJ, Hooper TD, et al. (2016) The application of the Global Trigger Tool: a systematic review. International Journal for Quality in Health Care 28: 640-649.
37. https://www.netflixprize.com/assets/ProgressPrize2007_KorBell.pdf
38. Hutchinson RA, Liu LP, Dietterich TG (2011) Incorporating Boosted Regression Trees into Ecological Latent Variable Models. In AAAI 11: 1343-1348.
39. Johnson R and Zhang T (2014) Learning nonlinear functions using regularized greedy forest. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(5): 942-954.
40. He X, Pan J, Jin O, et al. (2014) Practical lessons from predicting clicks on ads at facebook. In Proceedings of the Eighth International Workshop on Data Mining for Online Advertising. ACM: 1-9.
41. Nielsen D (2016) Tree Boosting with XGBoost-Why Does XGBoost Win "Every" Machine Learning Competition? Master's thesis, NTNU.