Natural Language Processing as a Predictor of Mortality in Intensive Care Unit Patients

Mario V. Fusaro (mvf1122@gmail.com)
Equum Research Institute
Christian Becker
Westchester Medical Center
Daniel Miller
Westchester Medical Center
Evan Finkel
Equum Research Institute
Ibrahim F. Hassan
Weill Cornell Medical College
Corey Scurlock
Equum Research Institute

Article

Keywords:

Posted Date: May 2nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1534066/v1

Abstract

Background: Artificial intelligence, specifically machine learning, has led to a series of medical publications in recent years. Several studies have used physiologic parameters with machine learning algorithms to derive robust clinical prediction tools. The goal of this study was to use natural language processing in conjunction with physiologic parameters to analyze clinical text data and create a clinical prediction tool.

Research Question: Can a natural language processing machine learning based clinical prediction tool accurately predict ICU mortality with relatively limited sample size?

Study Design and Methods: Three machine learning classifiers including a support vector machine (SVM), XGBoost Tree, and logistic regression were used to determine probability of hospital mortality in patients admitted to the intensive care unit (ICU) using physician notes and patient physiologic parameters. Discrimination for the binary outcome of life or death will be measured using receiver operating characteristic-area under the curve (AUC). Calibration will be assessed using chi-square goodness of fit.

Results: 7,555 patients were available for analysis with 5,288 being included in the derivation set and 2,267 in the validation set. Using the SVM algorithm, the AUC for hospital mortality was 0.895 (95% CI, 0.850-0.940). Calibration was also acceptable with chi square of 7.94 and p = 0.54. The XGBoost Tree algorithm resulted in the best discrimination for mortality with an AUC of 0.912 (95% CI, 0.881-0.943), but calibration was poor. Logistic regression resulted in an AUC of 0.868 (95% CI, 0.827-0.909) but calibration was also poor.

Conclusions: A support vector machine learning algorithm can use patient chart data, lab values, and physiologic parameters to generate a clinical prediction tool to predict mortality in ICU patients with a relatively limited dataset. For this study, SVM was superior to logistic regression which is used in many traditional ICU risk predictors.

Background

As Artificial Intelligence (AI) continues to develop, its uses and applications in medicine are rapidly evolving. AI is an umbrella term to describe several methods for processing information[1]. Machine learning, a form of AI, useful for predicting outcomes, has led to a series of medical publications in recent years[2-4]. The ability to collect vast amounts of data via electronic medical records (EMR) combined with advanced computer processing has facilitated the process[5]. Although machine learning can solve multiple problems including regression, clustering, and reinforcement learning, classification of an outcome is one of the most common medical uses. Machine learning classifiers could be helpful in managing the COVID-19 pandemic by predicting surge capacity for a geographically heterogeneous

intensive care unit (ICU) workforce or directing therapeutics to patients most likely to derive a benefit in resource depleted hospitals given variable patient response[6-9].

Machine learning classification uses one or many independent variables to assign a class or label to a dependent variable such as death or survival in the case of a binary predictor[10]. Similar to a clinical prediction tool, a set of examples (patients) will derive a rule or solution which will be tested on a validation group. Propensity score matching is also similarly constructed to a machine learning classifier[11]. In Intensive Care Unit (ICU) medicine, the most common algorithms used to make predictions are linear and logistic regression These are the basis for Acute Physiology and Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS), and Sequential Organ Failure Assessment (SOFA) scoring. Other algorithms including support vector machine (SVM), decision trees, and ensembles of multiple algorithms using the same independent variables can be used to derive this decision rule[4]. APACHE, SOFA, and SAPS are excellent predictors of mortality but exclude potentially useful information including radiology, pharmacy, and clinical impression. In order to capture textual information contained within the chart and convert this to a useful predictive information, natural language processing (NLP) can be used.

NLP is a machine learning technique in which a body of text (corpus) is deconstructed into single or multiple word fractions to be used as independent variables (tokens) for prediction of some dependent variable. As in logistic or linear regression, the words are converted datapoints for which to make some prediction, such as mortality. NLP in conjunction with basic physiologic parameters have recently been used to devise an ICU risk prediction system with very good performance[3]. These authors utilized a database of over 100,000 patients and logistic regression to arrive at the risk predictor.

Machine learning classification performance tends to improve with more examples available in derivation of the clinical prediction rule. As such, large, organized databases are typically necessary to arrive at meaningful results. More commonly, only smaller amounts of patient data are available for analysis, hindering progress. In this study, we have constructed a hospital risk predictor using limited physiologic variables and chart notes using XGBoost tree, logistic regression, and SVM algorithms in a relatively limited dataset of ICU patients.

Availability of Data and Materials:

The **datasets generated and analyzed in the current study are** from the publicly available Medical Information Mart for Intensive Care – III (MIMIC-III) database[5]. This de-identified relational database houses physiologic parameters and clinical data for over 50,000 patients dating from 2001-2012 from Beth Israel Deaconess Hospital in collaboration with Massachusetts Institute of Technology. Both organizations' institutional review board (IRB) approved waiver of consent for this deidentified dataset[12]. The New York Medical College also deemed this work to be IRB exempt. This study was carried out in accordance with the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement[13]**. All methods were performed in accordance with the relevant guidelines**

and regulations. The specific code with details of hyperparameter tuning are listed on github (https://github.com/Cobritra/NLPpublication.git).

Study Design And Methods

The patients included in this study were >16 years of age and required physiologic data from six hours before to 30 hours following ICU admission and included at least one physician note. The types and frequency of notes are listed in the appendix. This study only includes the first ICU admission of a hospital stay. The primary outcome was hospital mortality discrimination determined by receiver operating characteristic – area under the curve (AUC). A secondary outcome was test calibration. No blinding mechanism was used to measure outcome.

When constructing the prediction model, the independent variables used were clinical notes, physiologic variables, Glasgow Coma Scale (GCS) and pre-ICU Length of Stay (LOS). The first three notes ranging from 36 hours before and after ICU admission were used for NLP depending on availability. The note types included all available history and physical as well as consult notes. The initial lab values used were the average of the first three recorded during the prespecified timeframe: albumin, anion gap, creatinine, white blood cell (WBC) count, prothrombin time, hemoglobin, platelet count, serum sodium concentration, serum glucose, and serum potassium. The initial model also contained the average of the first three recorded: mean arterial pressure (MAP), systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, and temperature. The minimum recorded temperature and WBC count were also recorded. Additional variables included GCS and pre-ICU LOS. The final model excluded MAP, and minimum temperature because of redundancy within the model.

In order to use the clinical charts to make predictions, several data preprocessing steps were involved. The text was stripped of numbers, special characters, and certain words ('stop words') as they were quite common and do not add predictive value as listed in the appendix. After removal, the remaining text was transformed into single word strings (tokens) and converted to a variable (feature) matrix limited to 4000 tokens using count vectorization and label encoding with total counts per feature. A sample matrix displaying patient number on the y-axis and the frequency of each recorded word is displayed in the appendix and transformed by normalization. A matrix containing all non-text variables were merged with the text variable matrix for each respective patient. No blinding of predictor variables was used and the study size was determined by using the maximum available patients meeting inclusion criteria in the MIMIC-III database.

Missing clinical variables were imputed by taking the average value for that respective predictor variable across the full dataset and applying it to missing values, when necessary. Clinical prediction rules were derived using XGboost tree, SVM with linear kernel, and logistic regression with L2 regularization. Because of differences in the algorithms, certain transformations and data preprocessing steps were necessary for some algorithms but not others. No imputation was needed for the XGBoost tree

algorithm. The SVM algorithm required the matrix be normalized from 0 to 1. Feature importance for SVM was determined by classifier coefficient.

All calculations were performed using Python with the Anaconda© distribution. For the primary analysis, hospital mortality discriminatory performance was calculated using AUC while calibration was determined by chi square goodness of fit with an alpha level of 0.05. Normally distributed predictor variables were described using t-test, binary variables using chi square, and non-parametric continuous variables with Mann-Whitney test using IBM® SPSS® Version 25 software. For the primary outcome, the dataset was analyzed by k-folds with 10 partitions. The prediction tool was also analyzed using a random split into 70% derivation and 30% validation by computer sequence as part of sensitivity analysis.

Results

From 58,976 available patient examples, 7,555 patients with necessary notes and physiologic parameters were available for analysis (**Appendix**). The derivation set included 5,288 patients while the validation set was comprised of 2,267 patients. The average age was 64 (Interquartile Range [IQR], 52.6-78.7), hospital mortality rate was 8.8%, and the population was 45% female. Additional clinical features are listed in **(Table 1)**. The most common note titles were '*Intensivist Notes*', '*Physician Resident Admission Notes*', '*Physician Attending Admission Note – MICU*, and '*Physician Attending Progress Note*' and are listed in the appendix. The period over which this data set is collected was 2001-2012 however exact dates are censored. Of all the variables retrieved 18.3% were missing.

Using the SVM learning algorithm with text only, discrimination for the outcome of mortality displayed an AUC of 0.871. The AUC for clinical variables only was 0.743 (95% CI, 0.709-0.777). When both types of variables were combined the AUC was 0.891 using the 70/30 split (**Figure 1**) and then 0.895 (95% CI, 0.850-0.940) when using k-folds with 10 partitions which was significantly different than the value with clinical variables alone ($p = 0.022$). Calibration by using chi-square goodness of fit was $\chi^2 = 7.94$ and $p = 0.54$ (**Table 2**). The XGBoost Tree algorithm resulted in discrimination for mortality with AUC of 0.912 (95% CI, 0.881-0.943) using k-folds with 10 partitions. Calibration revealed $\chi^2 = 857$ and $p < 0.01$ for this model (appendix). Discrimination using logistic regression for this model resulted in an AUC of 0.868 (95% CI, 0.827-0.909) and $\chi^2 = 379$ and $p < 0.01$ for calibration (**Appendix**).

Discussion

Our study found that SVM machine learning algorithms using natural language processing are capable of generating acceptable hospital mortality predictions for ICU patients ICU patients with AUC and calibration similar to what was noted in the description of APACHE-IV by Zimmerman et al (**Figure 1**)[6]. The XGboost algorithm was able to provide higher discrimination compared with SVM however calibration was poor for this method. The logistic regression model using this data set for mortality prediction had slightly lower power to discriminate and poorer calibration. APACHE-IV and Marafino et al

utilized logistic regression with patient numbers in excess of 100,000 to achieve similar discrimination and calibration[3,14]. It is possible that SVM model has both excellent discrimination and calibration however it is also possible that this model suffers from overfitting. This algorithm will need to be tested against an external validation set to confirm the findings. A learning curve displaying the SVM model suggests bias, variance and possibly overfitting is muted and the model would overall benefit from an increase in examples (**Appendix**).

Although the terms artificial intelligence and machine learning are being used more regularly with regard to medical innovation, these basic methods have been used for decades. Classic machine learning algorithms include linear and logistic regression which in the past has been the basis for many clinical prediction tools. The basic structure of a machine learning classifier is that of a clinical prediction tool with a derivation data set and validation data set with a solution generated in between. As more of the machine learning repertoire is utilized in medical decision making, newer and possibly better predictive models could be devised. Accuracy of clinical prediction tools might be limited using traditional independent variables such as vital signs and physiologic parameters. Addition of information sources from the written chart can introduce a new set of variables from which to make medical decisions. Natural language processing allows for a model to integrate this additional information. Previously, ICU mortality predictions had been made using natural language processing and logistic regression[3]. Exploration of additional machine learning techniques such as neural networks may even yield better performance with adequate available data. In this instance, the SVM algorithm outperformed logistic regression and had better calibration than XGBoost. Determining the optimal algorithm can be difficult and so deriving multiple models using several algorithms is necessary. SVM algorithms tend to perform well in high dimensional spaces such as this model with >4,000 independent variables. The XGBoost algorithm would tend to have better discrimination overall however this model has significant discrepancies when comparing observed to predicted outcomes across the intermediate deciles of risk (**Appendix**). The SVM algorithm was able to maintain accuracy throughout all deciles of risk (**Figure 2**).

Several limitations exist in this study including a single data source, potential issues with generalizability of findings, limited sample size, and speed of generating SVM predictions with larger datasets. Although this study was carried out on data from a single source with limited numbers, these issues are characteristic of hospitals with limited datasets. This may limit the model's external validity. Marafino et al also noted geographical differences in performance possibly due to idiosyncrasies in documentation[3]. In examining the feature importance (independent variable), several words are highly predictive of poor outcomes including: 'unresponsive', 'dnr', 'prognosis', 'family' and 'comfort'. Although this is not surprising, individual differences in documentation between providers could lead to variable performance. However, if the purpose of the prediction is internal benchmarking, risk prediction, or single center triage, then this becomes less of a limitation. In this group of models, the SVM algorithm had both good discrimination and calibration however was the slowest to run. As examples become more plentiful, SVM algorithms can take longer to derive a solution. Another limitation with this dataset might be the timeframe from which it was collected. Over the timeframe during which this data was collected, critical

care practice patterns may have changed leading to improved outcomes which might affect prediction accuracy when comparing a more recent dataset[15].

AI based tools could be transformational when used in an ICU setting to improve accuracy of prediction and prognostication models, such as for sepsis and ICU mortality. Basic machine learning methods have given way for a greater role and wider gamut of functions within medicine. Recent studies have developed algorithms that incorporate multiple variables to help resemble clinician decision making, reduce inter-clinician variability, and even make diagnosis and treatment decisions[16,17]. Specific use cases with regard to the COVID-19 pandemic may include algorithms which predict which patients will derive the most benefit from novel and scarce therapeutics so as to best target limited resources. An additional potential use case might be a triage system using an NLP/physiology based outcome predictor to best target Tele-Critical care services in regions without adequate intensivist support[6]. In addition to evolving use cases, continual improvements in clinical prediction tool accuracy must be pursued to instill confidence in any algorithm used for patient care decision making.

Conclusions

A support vector machine learning algorithm can use patient chart data as well as lab and physiologic parameters to generate a hospital mortality clinical prediction tool for ICU patients with a relatively small dataset on par with APACHE-IV. Tools like these can be integrated with clinical workflows to help in decision making as well as help guide and scale the critical care workforce. Further exploration should focus on improving the accuracy of these predictions to better serve patients.

Abbreviations

Artificial Intelligence (AI), Intensive Care Unit (ICU), Severity of Illness (SOI), Length of Stay (LOS), Electronic Medical Record (EMR), Acute Physiology Score (APS), Acute Physiology and Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS),

Declarations

Conflict of Interest: The authors confirm to have read BioMed Central's guidance on competing interests. All authors declare not to have any competing interests in the manuscript.

Funding: None.

Take Home Message:

Natural Language Processing can be used in conjunction with physiologic variables to accurately predict ICU mortality.

Acknowledgements:

Author Contributions: M.F. had full access to the data and assume responsibility for the integrity and accuracy of data analysis. E.F., C.B., and C.S. contributed substantially to the study design, data analysis, interpretation, and writing of the manuscript. D.M. and I.H. contributed substantially to the data interpretation and writing of the manuscript.

Funding:

None

References

1. Amisha, Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. *J Family Med Prim Care*. 2019;8(7):2328-2331.
2. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK biobank participants. *PLoS One*. 2019;14(5):e0213653.
3. Marafino BJ, Park M, Davies JM, et al. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. *JAMA Netw Open*. 2018;1(8):e185097.
4. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the super ICU learner algorithm (SICULA): A population-based study. *Lancet Respir Med*. 2015;3(1):42-52.
5. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035.
6. Krouss M, Allison MG, Rios S, et al. Rapid implementation of telecritical care support during a pandemic: Lessons learned during the coronavirus disease 2020 surge in new york city. *Crit Care Explor*. 2020;2(11):e0271.
7. Beigel JH, Tomashek KM, Dodd LE, et al. Remdesivir for the treatment of covid-19 - final report. *N Engl J Med*. 2020.
8. RECOVERY Collaborative Group, Horby P, Lim WS, et al. Dexamethasone in hospitalized patients with covid-19 - preliminary report. *N Engl J Med*. 2020.
9. Angus DC, Kelley MA, Schmitz RJ, White A, Popovich J,Jr, Committee on Manpower for Pulmonary and Critical Care Societies (COMPACCS). Caring for the critically ill patient. current and projected workforce requirements for care of the critically ill and patients with pulmonary disease: Can we meet the requirements of an aging population? *JAMA*. 2000;284(21):2762-2770.
10. Saria S, Butte A, Sheikh A. Better medicine through machine learning: What's real, and what's artificial? *PLoS Med*. 2018;15(12):e1002721.
11. Westreich D, Lessler J, Funk MJ. Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin*

Epidemiol. 2010;63(8):826-833.

12. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter intelligent monitoring in intensive care II: A public-access intensive care unit database. *Crit Care Med*. 2011;39(5):952-960.
13. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63.
14. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute physiology and chronic health evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit Care Med*. 2006;34(5):1297-1310.
15. Santacruz CA, Pereira AJ, Celis E, Vincent JL. Which multicenter randomized controlled trials in critical care medicine have shown reduced mortality? A systematic review. *Crit Care Med*. 2019;47(12):1680-1691.
16. Amisha, Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. *J Family Med Prim Care*. 2019;8(7):2328-2331.
17. Lovejoy CA, Buch V, Maruthappu M. Artificial intelligence in the intensive care unit. *Crit Care*. 2019;23(1):7-018-2301-9.

Tables

Table 1: Median physiologic characteristics of the derivation and validation cohorts.

Variable	Derivation (n=5,288)	IQR	Validation (n=2,267)	IQR
Pre-ICU LOS (h)	0.0	5.9	0.0	6.0
Hospital Mortality (%)	9.0	-	8.3	-
Age (y)	65.9	25.8	64.5	27.1
GCS	14.7	2.7	14.7	3.0
Systolic BP (mmHg)	120	33.3	119	29.7
Diastolic BP (mmHg)	62	16.7	62	14.7
Heart Rate (BPM)	86	25.1	88	26.0
Respiratory Rate (RPM)	17	6.7	18	7.0
Temp ^{o}F	97.9	1.7	98.0	1.7
Albumin g/dl	3.1	1.0	3.1	1.0
Anion Gap mEq/L	13.0	4.0	13.0	4.0
Creatinine mg/dl	1.0	0.7	0.9	0.7
WBC 1000/µl	10.0	6.2	9.9	6.2
Lowest WBC 1000/µl	9.3	5.7	9.1	5.7
PT (s)	14.4	3.3	14.4	3.0
Hemoglobin g/dl	10.8	2.8	10.7	2.8
Platelet 1000/µl	214	117	213	122
Serum Sodium mEq/L	139	4.3	139	4.0
Serum Potassium mEq/L	4.1	0.7	4.1	0.7
Serum Glucose mg/dl	121	48.8	121	50.5

ICU = Intensive Care Unit, IQR = Interquartile Range, LOS= Length of Stay, GCS = Glasgow Coma Scale, BP = Blood Pressure, BPM = Breaths per Minute, RPM = Respirations Per Minute, Temp = Temperature in Fahrenheit, WBC = white blood cell count, PT = Prothrombin Time

Table 2: A comparison of risk deciles between observed and predicted mortality with model derived by support vector machine algorithm. Chi square = 7.94, p = 0.54, df = 9.

Risk Decile	Observed Deaths	%	Predicted Deaths	%	Difference (%)
1	1	0.4	1	0.2	0.2
2	0	0.0	1	0.6	-0.6
3	3	1.3	2	0.9	0.4
4	0	0.0	3	1.4	-1.4
5	3	1.3	5	2.2	-0.9
6	9	3.9	7	3.2	0.7
7	10	4.4	12	5.1	-0.7
8	22	9.7	18	7.9	1.8
9	36	15.9	33	14.4	1.5
10	105	46.5	105	46.7	-0.2

df = degrees of freedom

Figures

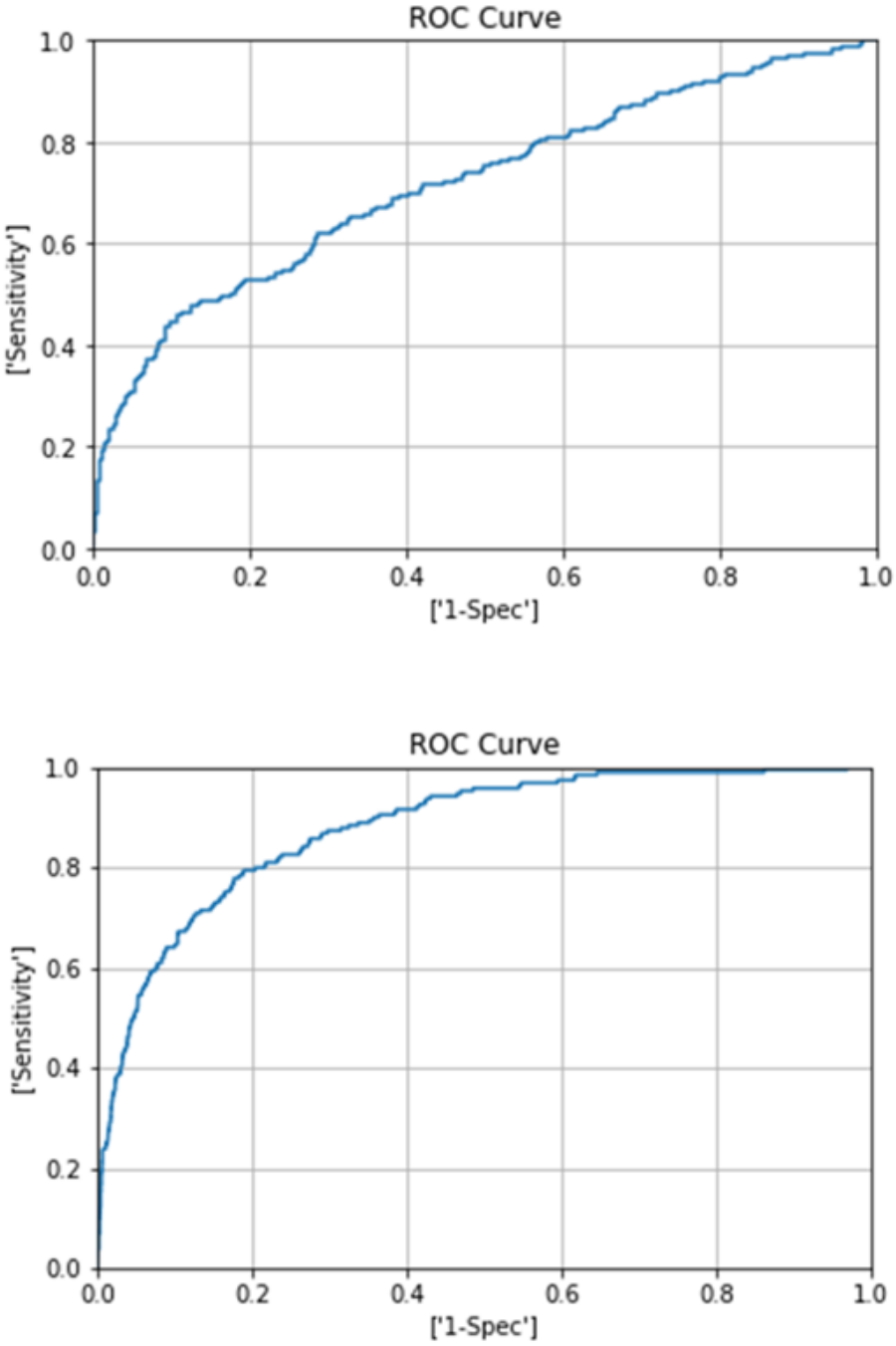

Figure 1

Receiver Operating Characteristic – Area Under the Curve (AUC) for hospital mortality model derived from support vector machine algorithm with 70/30 split. Top: AUC of model with clinical parameters only, 0.717. Bottom: AUC of model using both clinical variables and natural language processing, 0.891.

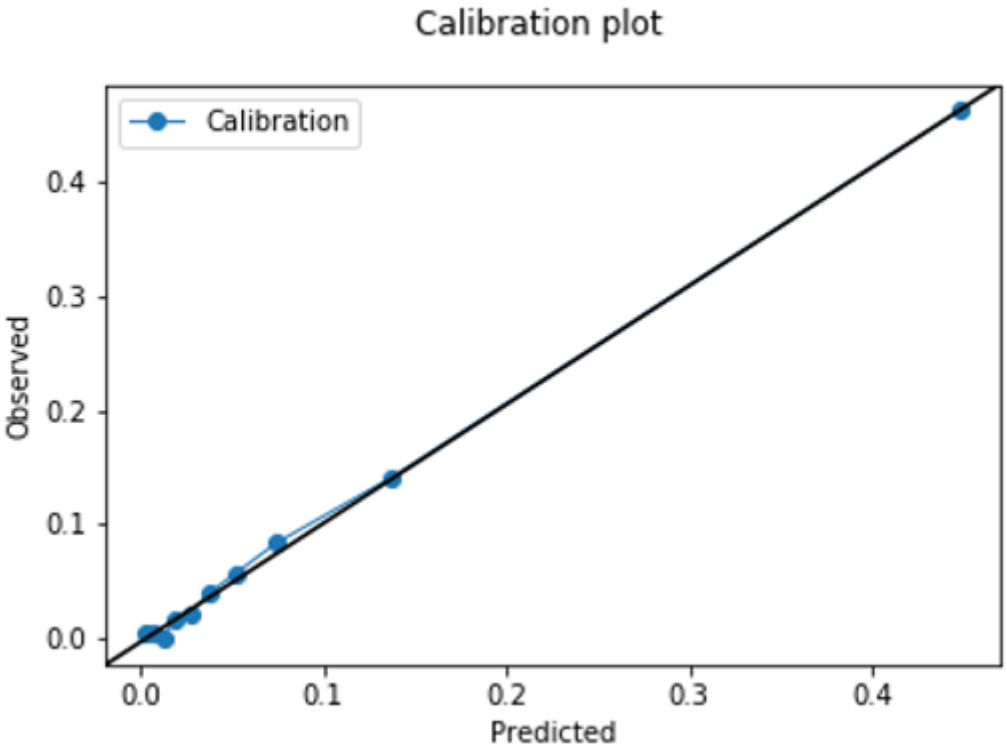

Figure 2

Comparison of observed compared to predicted mortality incidence proportion in validation cohort using the support vector machine algorithm for hospital mortality.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NLPAppendix.docx