big data and cognitive computing

MDPI

Article

A Simple Free-Text-like Method for Extracting Semi-Structured Data from Electronic Health Records: Exemplified in Prediction of In-Hospital Mortality

Eyal Klang [1,*,†], Matthew A. Levin [2], Shelly Soffer [3], Alexis Zebrowski [4], Benjamin S. Glicksberg [5], Brendan G. Carr [4], Jolion Mcgreevy [4], David L. Reich [2] and Robert Freeman [6]

[1] Chaim Sheba Medical Center, Department of Diagnostic Imaging, Affiliated to Tel-Aviv University, Tel Aviv-Yafo 52621, Israel

[2] Department of Anesthesiology, Perioperative and Pain Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; Matthew.Levin@mssm.edu (M.A.L.); david.reich@mountsinai.org (D.L.R.)

[3] Internal Medicine B, Assuta Medical Center, Ben-Gurion University of the Negev, Be'er Sheva 7747629, Israel; soffer.shelly@gmail.com

[4] Department of Emergency Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; Alexis.Zebrowski@mountsinai.org (A.Z.); Brendan.Carr@mountsinai.org (B.G.C.); Jolion.Mcgreevy@mountsinai.org (J.M.)

[5] Hasso Plattner Institute for Digital Health at Mount Sinai, New York, NY 10065, USA; benjamin.glicksberg@mssm.edu

[6] Institute for Healthcare Delivery Science, Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; Robert.Freeman@mountsinai.org

* Correspondence: eyal.klang@mountsinai.org; Tel.: +1-972-306-8504

† Address: Sheba Medical Center Hospital-Tel, Hashomer, Ramat Gan 52621, Israel.

check for updates

Citation: Klang, E.; Levin, M.A.; Soffer, S.; Zebrowski, A.; Glicksberg, B.S.; Carr, B.G.; Mcgreevy, J.; Reich, D.L.; Freeman, R. A Simple Free-Text-like Method for Extracting Semi-Structured Data from Electronic Health Records: Exemplified in Prediction of In-Hospital Mortality. *Big Data Cogn. Comput.* **2021**, *5*, 40. https://doi.org/10.3390/bdcc5030040

Academic Editor: Min Chen

Received: 17 June 2021
Accepted: 26 August 2021
Published: 29 August 2021

Abstract: The Epic electronic health record (EHR) is a commonly used EHR in the United States. This EHR contain large semi-structured "flowsheet" fields. Flowsheet fields lack a well-defined data dictionary and are unique to each site. We evaluated a simple free-text-like method to extract these data. As a use case, we demonstrate this method in predicting mortality during emergency department (ED) triage. We retrieved demographic and clinical data for ED visits from the Epic EHR (1/2014–12/2018). Data included structured, semi-structured flowsheet records and free-text notes. The study outcome was in-hospital death within 48 h. Most of the data were coded using a free-text-like Bag-of-Words (BoW) approach. Two machine-learning models were trained: gradient boosting and logistic regression. Term frequency-inverse document frequency was employed in the logistic regression model (LR-tf-idf). An ensemble of LR-tf-idf and gradient boosting was evaluated. Models were trained on years 2014–2017 and tested on year 2018. Among 412,859 visits, the 48-h mortality rate was 0.2%. LR-tf-idf showed AUC 0.98 (95% CI: 0.98–0.99). Gradient boosting showed AUC 0.97 (95% CI: 0.96–0.99). An ensemble of both showed AUC 0.99 (95% CI: 0.98–0.99). In conclusion, a free-text-like approach can be useful for extracting knowledge from large amounts of complex semi-structured EHR data.

Keywords: electronic health records; machine learning; gradient boosting

1. Introduction

In the last decade, the medical world has been exposed to two important concepts related to digital information: Big-Data and Artificial Intelligence (AI) [1]. Bringing these two concepts together enables the creation of increasingly accurate prediction models.

One setting that can benefit from decision support tools is the emergency department (ED) triage. EDs are becoming increasingly crowded, impairing patient outcomes [2–6]. Decision support tools can aid the expedited assessment of patients during initial ED triage. Several clinical triage acuity scores have been developed. The most commonly used is the five-level emergency severity index (ESI). In recent years, studies have evaluated different decision support tools in the triage [7–9]. Yet, most of these studies used a discrete number of variables.

Today, the electronic health record (EHR) stores a wealth of information for each patient, both as tabular data and as free text. Patient cohort data is usually stored in a two axes matrix. The x-axis (rows) represents individual patients, and the y-axis (columns) represents data for each patient. While many machine-learning models strive to use a large number of rows, usually a limited number of columns are utilized.

The Epic EHR (Epic Systems Corporation, Verona, WI) is one of the most commonly used EHR in the United States. It is estimated that more than 250 million patients have a current electronic record in Epic [10]. Epic stores a majority of the data inside documents or structures called "flowsheets". These fields contain vast amount of semi-structured items that pertain to patient assessment. Flowsheet data lack a well-defined external ontology or data dictionary and are often unique to each implementation of Epic. This makes utilizing the information contained within them, which may include valuable clinical observations, quite difficult. We hypothesized that a free-text approach could help utilize the semi-structured Epic data.

We evaluated a simple free-text-like method to extract semi-structured EHR data. We tested this method on two machine learning models and on an ensemble of both. First, we trained a logistic regression model. Logistic regression is a well-established model. The model is easy to implement and easy to interpret and does not require significant resources. Second, we trained the XGboost implementation of gradient boosting algorithm. Gradient boosting is a machine learning algorithm where multiple weak learners are trained to augment each-other and together produce superior results. At each stage a new decision tree is learned with the aim to correct errors made by existing trees. As a non-linear method, it often out-performs linear models, when higher order relationships exist in the data. Gradient boosting has also surpassed other machine learning algorithms in a number of data challenges.

As a use case we demonstrate using this method in predicting in-hospital mortality during ED triage.

2. Materials and Methods

The Mount Sinai Hospital institutional review board (IRB) approval was granted for this retrospective study. Informed consent was waived by the IRB committee.

The study was conducted at the Mount Sinai Hospital (MSH) New York City. This is a large academic tertiary center with approximately 110,000 annual ED visits. The study's time frame was between 1 January 2014, to 31 December 2018.

We retrieved records of consecutive adult (age $\geq$ 18) patients admitted to the ED. Erroneously created and duplicate charts were excluded. We also excluded visits without triage notes and patients who died within 30 min from the triage note.

Both structured and free-text time-stamped data were retrieved from the EHR. All items were limited to those documented up to 30 min from triage note. Data points include Demographics: age, sex, ethnicity; Arrival mode (walk-in, by ambulance, or by intensive care ambulance); Chief complaints; Comorbidities, coded as International Classification of Diseases (ICD-10) and grouped using the diagnostic clinical classification software (CCS); First vital signs measurements; Acuity level (ESI); Laboratory orders; Nursing and physician text notes (free-text); and all Epic's flowsheet records from the visit.

The primary outcome was in-hospital death within 48 h. As a secondary outcome we evaluated overall in-hospital death.

2.1. Data Representation

Both semi-structured and free-text were encoded using a Bag of Words (BoW) approach [11]. BoW is a commonly used approach in natural language processing (NLP). In BoW, a text paragraph is represented as an unordered collection (bag) of its words. A classifier classifies the paragraphs based on the frequency of words in the "bags". Sparse matrix representation was used for the BoW collections.

BoW collections were used to represent the following data: nursing and physician free-text notes, flowsheet records, comorbidities, chief complaints and lab orders. For each of these items we also encoded the time in minutes from triage note to the item as a separate BoW collection.

For the flowsheet field, BoW containers were encoded in three ways: (1) type (e.g, "ED_physical"); (2) item (e.g, "Chest_auscultation"); (3) item + value (e.g, "Chest_auscultation:rales").

We also created a BoW container to represent "past stories". This encoded the number of previous ED visits and hospitalizations, number of days to previous visits, type of ward if hospitalized and chief complaints during the previous visits.

All other variables (demographics, mode of arrival, vital signs) were concatenated to the BoW collections.

2.2. Machine Learning Models

Two machine-learning methods were trained: gradient boosting and logistic regression. We tested logistic regression with term frequency-inverse document frequency (LR-tf-idf) [11]. An ensemble of LR-tf-idf and gradient boosting was also evaluated. Figure 1 presents the schematics of the models.

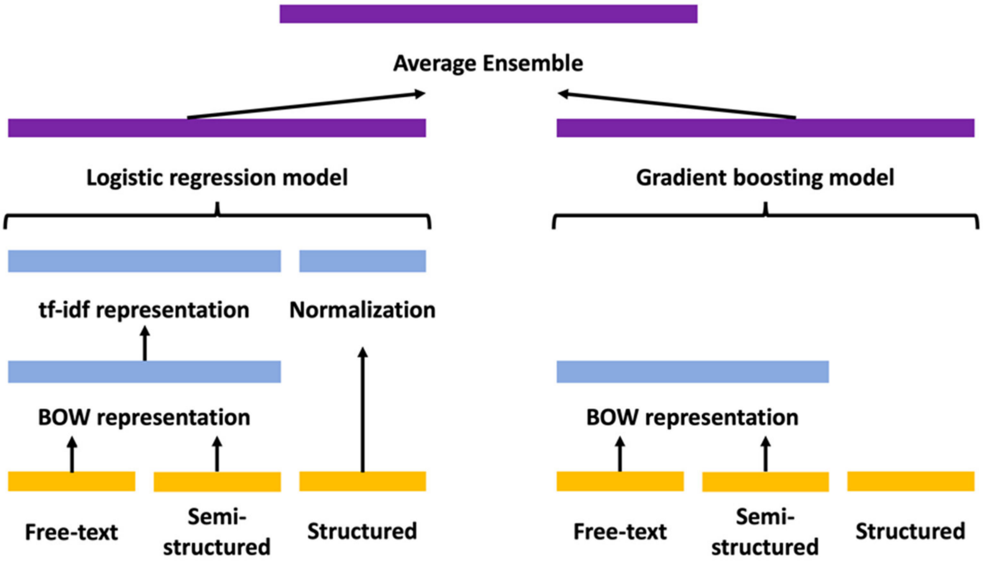


Figure 1. Schematics of the models' design.

Continuous variables were normalized (Z-scores) for the logistic regression model. Normalization was not used for the gradient boosting model. As this model "cuts" above and below the desired value. Thus, it is not affected by linear transformations.

Models were trained on data from the years 2014–2017 and tested on data from the year 2018. This ensures no chronological leakage of information.

2.2.1. Logistic Regression

A term frequency-inverse document frequency (*tf-idf*) approach was employed to the BoW collections. *Tf-idf* balances the importance of a word to the document (*tf*) and the frequency of the word in the corpus (*idf*).

The *tf-idf* formula for each word (*w*) in one document is:

$$wscore = tf * idf$$

$$tf = \frac{Number\ of\ w\ in\ the\ document}{Total\ number\ of\ words\ in\ the\ document}$$

$$idf = \log \frac{Total\ number\ of\ documents}{Number\ of\ documents\ containing\ w}$$

The logistic regression hyperparameters included default l2 regularization with C = 1.0, and number of iterations = 2000. Variables with missing data were not included in the logistic regression model as experimentation with imputations did not show benefit. Data balancing was not used for the logistic regression as it did not improve the results.

2.2.2. Gradient Boosting

We used the XGBoost implementation of the gradient boosting algorithm [12]. This model uses multiple tree-based classifiers trained to correct errors made by the previous trees. The default hyper-parameters were used for the model: eta = 0.3, max depth = 3. We set n_estimators = 1000. Imputations of missing values were handled by the XGBoost model. Scale balancing of the XGBoost was set to the default scale pos weight = 1. since weight balancing did not affect the gradient boosting model, it was not applied.

2.2.3. Ensemble

Ensemble averaging is the process of averaging of multiple models' predictions to improve the desired output. This process is opposed to creating just one model. The ensemble of several models frequently performs better than any individual model's predictions, since the errors of the models "average out." We evaluated an ensemble averaging of the LR-tf-idf and gradient boosting outputs.

2.3. Statistical Analysis

All development and statistical analysis were carried out using Python (Version 3.6.5).

Continuous variables are reported as the median with the spread reported as the Interquartile range (IQR). Categorical variables are reported as percentages. Continuous variables were compared using 1-way analysis of variance (ANOVA). Categorical variables were compared using the chi-square test. The Area under the curve (AUC) metric was used to compare model performance on the testing data (the year 2018).

To analyze the importance of single terms/words in the flowsheet and in the free-text BoW collections, we used the mutual information formula [13]. This formula measures the joint mutual information between the mortality class (C) and the term/word (*W*):

$$Mutual\ Information = \sum\sum P(C,\ W) * Log\ \frac{P(C,W)}{P(C)P(W)}$$

Youden's index was used to find an optimal sensitivity-specificity cutoff point on the receiver operating characteristic (ROC) curve. Sensitivity, specificity, false-positive rate (FPR), negative predictive value (NPV), positive predictive value (PPV) and F1-score were also evaluated for fixed specificities of 90% and 99%. Bootstrapping validations (1000 bootstrap resamples) were used to calculate 95% confidence intervals (CI) for all metrics.

3. Results

3.1. Study Cohort

During the five-year study period, the MSH recorded 546,186 ED visits. After exclusion, the cohort consisted of 412,901 ED visits (Figure 2). Overall, 2803 in-hospital mortality cases (0.7%) were identified. Of them, 703 (0.2%) died within 48 h of ED admission. The median time to death was 7 days (IQR: 2–16 days). 42 patients died within 30 min from triage note and were excluded.

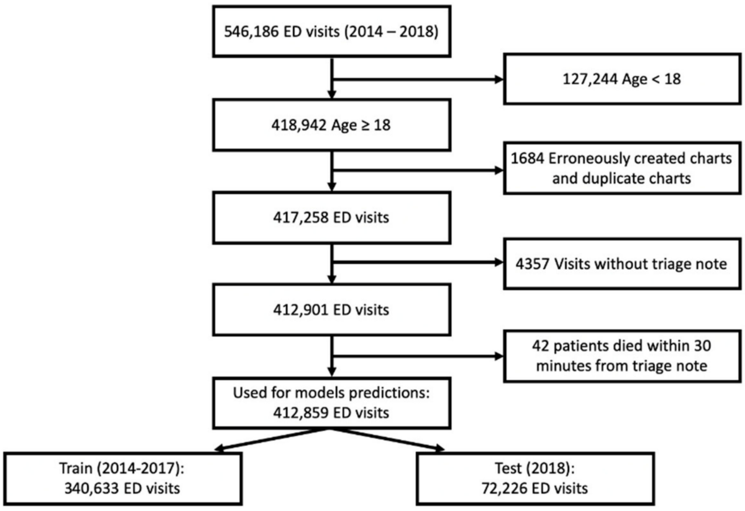

546,186 ED visits (2014 – 2018)
127,244 Age < 18
418,942 Age ≥ 18
1684 Erroneously created charts and duplicate charts
417,258 ED visits
4357 Visits without triage note
412,901 ED visits
42 patients died within 30 minutes from triage note
Used for models predictions: 412,859 ED visits
Train (2014-2017): 340,633 ED visits
Test (2018): 72,226 ED visits

Figure 2. Inclusion flow chart.

Patient characteristics of both the training and testing dataset are presented in Table 1. Significant differences were observed between patients who died and those who survived (Table 1). Of note, about half of the mortality cases had known cardiovascular and oncological diseases. Table 2 describes how the data were distributed across the training and testing datasets.

Table 1. Patients' characteristics.

	Survived (*n* = 410,098, 99.3%)	**48-h In-Hospital Mortality** (*n* = 703, 0.2%)	**In-Hospital Mortality after 48 h** (*n* = 2100, 0.5%)	***p* Value**
		Demographics		
Age, median (IQR), y	48.0 (30.0–63.0)	74.0 (62.0–85.0)	68.0 (59.0–81.0)	<0.001
Male, N. (%)	239,249 (58.3)	369 (52.5)	1017 (48.4)	<0.001
Black, N (%)	128,655 (31.4)	185 (26.3)	699 (24.9)	<0.001
White, N (%)	77,216 (18.8)	173 (24.6)	869 (31.0)	<0.001
		Hospital course		
LOS, median (IQR), Hours	4.0 (2.0–17.0)	12.0 (4.0–27.0)	271.0 (143.8–494.0)	<0.001
ICU, N. (%)	6752 (1.6)	185 (26.3)	1038 (49.4)	<0.001
Death in ED, N. (%)	0	357 (50.8)	24 (1.1)	<0.001
		Vital signs		
SBP, median (IQR), mmHg	132.0 (119.0–149.0)	113.0 (93.0–138.0)	122.0 (106.0–142.0)	<0.001
DBP, median (IQR), mmHg	73.0 (65.0–83.0)	61.0 (53.0–76.0)	66.0 (57.0–77.0)	<0.001
Heart rate, median (IQR), b/min	84.0 (74.0–96.0)	93.0 (71.5–117.0)	95.0 (79.0–111.0)	<0.001
Temperature, median (IQR), Fahrenheit	97.5 (96.8–98.2)	97.2 (96.3–98.2)	97.5 (96.8–98.6)	<0.001
Respirations, median (IQR), N/min	18.0 (18.0–20.0)	20.0 (18.0–24.0)	20.0 (18.0–20.0)	<0.001
O2 saturation, median (IQR), %	98.0 (97.0–99.0)	96.0 (91.0–99.0)	97.0 (95.0–99.0)	<0.001
		Accumulated data		
Physician text, N. (%)	101,050 (24.6)	375 (53.3)	773 (36.8)	<0.001
Number of free-text words, median (IQR), N.	25.0 (14.0–71.0)	94.0 (23.0–484.0)	35.0 (18.0–246.5)	<0.001
Number of flowsheet records, median (IQR), N.	34.0 (25.0–45.0)	41.0 (26.0–61.0)	38.0 (28.0–54.0)	<0.001
		Comorbidities		
CVD, N. (%)	113,509 (27.7)	365 (51.9)	1082 (51.5)	<0.001
DM, N. (%)	101,087 (24.6)	267 (38.0)	772 (36.8)	<0.001

Table 1. *Cont.*

	Survived (*n* = 410,098, 99.3%)	48-h In-Hospital Mortality (*n* = 703, 0.2%)	In-Hospital Mortality after 48 h (*n* = 2100, 0.5%)	*p* Value
HTN, N. (%)	55,008 (13.4)	116 (16.5)	359 (17.1)	<0.001
COPD, N. (%)	22,774 (5.6)	113 (16.1)	333 (15.9)	<0.001
Cancer, N. (%)	77,098 (18.8)	244 (34.7)	988 (47.0)	<0.001
Number of comorbidities, median (IQR), N.	2 (0–7)	3 (0–10)	3 (0–9)	<0.001

Abbreviations: IQR interquartile range; LOS length of stay; ED emergency department; SBP systolic blood pressure; DBP diastolic blood pressure; CVD cardiovascular disease; DM diabetes mellitus; HTN hypertension; COPD chronic obstructive pulmonary disease.

Table 2. Data distribution across the training and testing datasets.

	Overall	Survived	In-Hospital Mortality within 48 h	In-Hospital Mortality after 48 h
Train	340,633	338,325 (99.3%)	540 (0.2%)	1768 (0.5%)
Test	72,226	71,773 (99.3%)	121 (0.2%)	332 (0.5%)

3.2. Data Analysis

Figure 3 presents distribution of the sizes of different data types. Semi-structured flowsheet data was the largest, followed by free-text and lastly, structured data. On average, each patient had 37.9 (±18.4) flowsheet records accumulated within 30 min from triage.

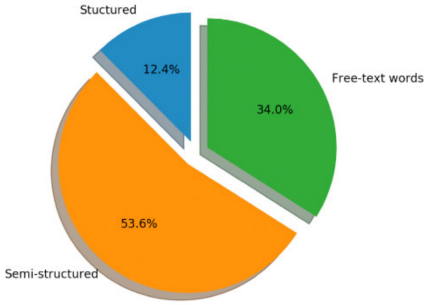


Figure 3. Distribution of the sizes of different data types.

Flowsheet terms combinations associated with 48-h mortality include: “acuity level: 1”, and terms related to mechanical ventilation and sepsis (Table A1 in Appendix A). Terms associated with overall in-hospital mortality include: “acuity level: 1”, “acuity level: 2”, “sepsis” and “supine position” (Table A1).

Free-text words associated with 48-h mortality include resuscitation measures such as “CPR”, “ACLS”, “arrest” and “intubation” (Table A2). Free-text words associated with overall mortality include words related to transfer and disposition such as “EMS”, “from”, “admission” and also words related to resuscitation such as “CPR”, and “arrest” (Table A2).

3.3. Machine-Learning Models

Adding tf-idf markedly improved the logistic regression model (Table 3). For 48-h mortality, the AUC rose from 0.92 to 0.98. For overall mortality, the AUC rose from 0.86 to 0.95.

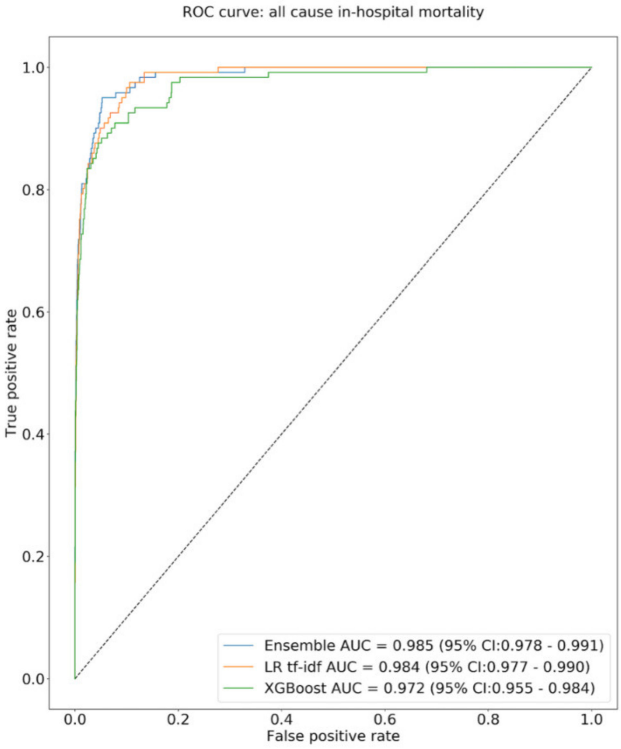
The gradient boosting model showed an AUC 0.97 (95% CI: 0.96–0.99) for 48-h mortality and 0.95 (95% CI: 0.95–0.96) for overall mortality.

The ensemble model showed only a small increase over single models. For 48-h mortality, the AUC increased to 0.99 (95% CI: 0.98–0.99) and for overall mortality to 0.96 (95% CI: 0.98–0.99). Yet, the ensemble did show better results compared to the gradient boosting for predicting 48-h mortality and better results compared to logistic regression for predicting overall mortality (Figure 4a,b). Calibration plots of the models are presented in Figure 5a,b. Figure 6a,b show the Precision-Recall (PR) curve for in hospital mortality within 48 h and overall in-hospital mortality. Figure 7a,b demonstrate the confusion matrix of the ensemble model predictions for in hospital mortality within 48 h and overall in-hospital mortality using Youden's index cut-off.

Table 3. Results of the logistic regression model. Comparison of predictions with and without employing tf-idf.

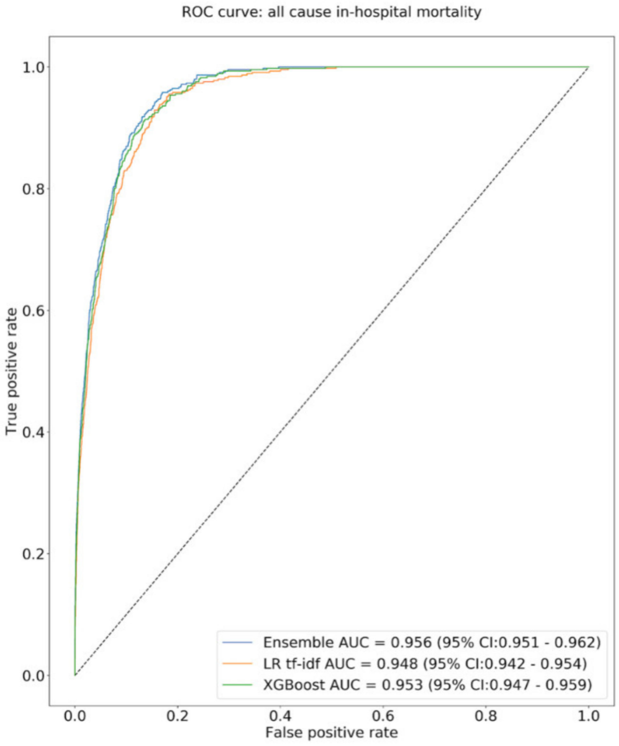
	Logistic Regression	**Logistic Regression with Tf-Idf**
AUC: Mortality within 48 h	0.92 (95% CI: 0.89–0.95)	0.98 (95% CI: 0.98–0.99)
AUC: Overall in-hospital mortality	0.86 (95% CI: 0.84–0.88)	0.95 (95% CI: 0.94–0.95)

For 48-h mortality, the ensemble showed sensitivity 95%, specificity 95% (FPR 1:20) and PPV 0.03 (Tables 4 and 5). For FPR 1:100 the ensemble showed sensitivity 74% and PPV 0.11. Figure 8a,b present word clouds of terms importance.



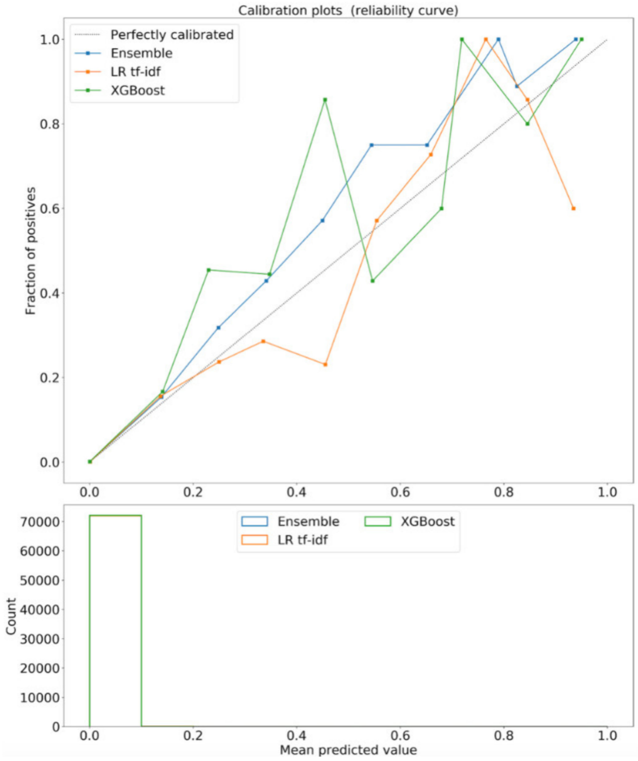
(**a**)

Figure 4. *Cont.*



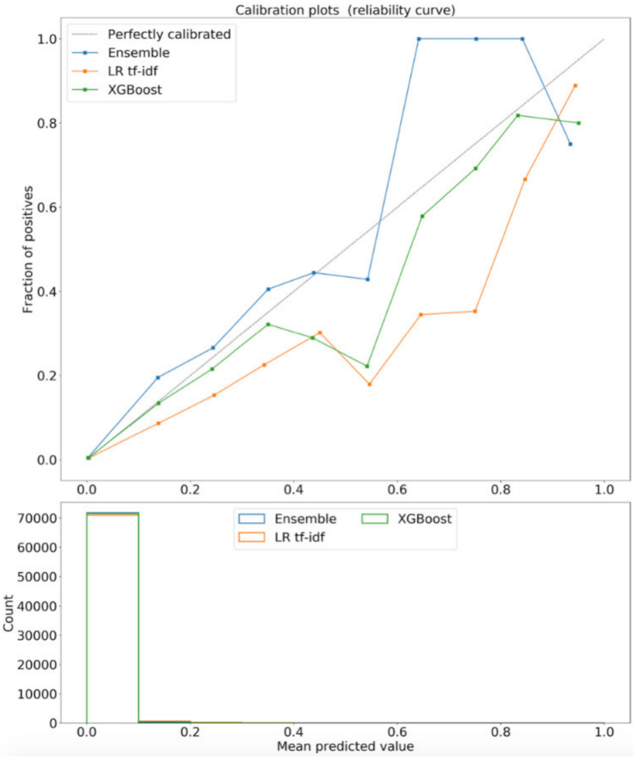
(**b**)

Figure 4. (**a**) Receiver operator curves (ROC) for predicting 48-h mortality (**b**) ROC for predicting overall in-hospital mortality.



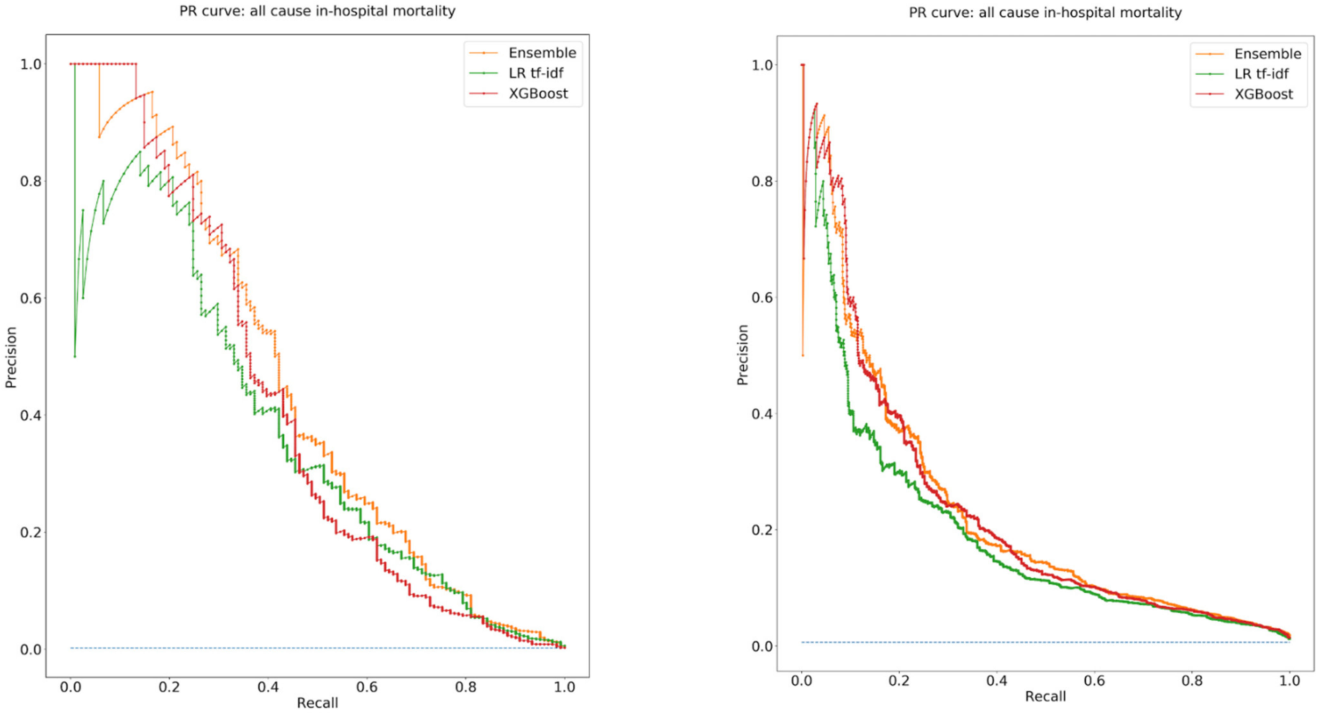
(**a**)

Figure 5. *Cont.*



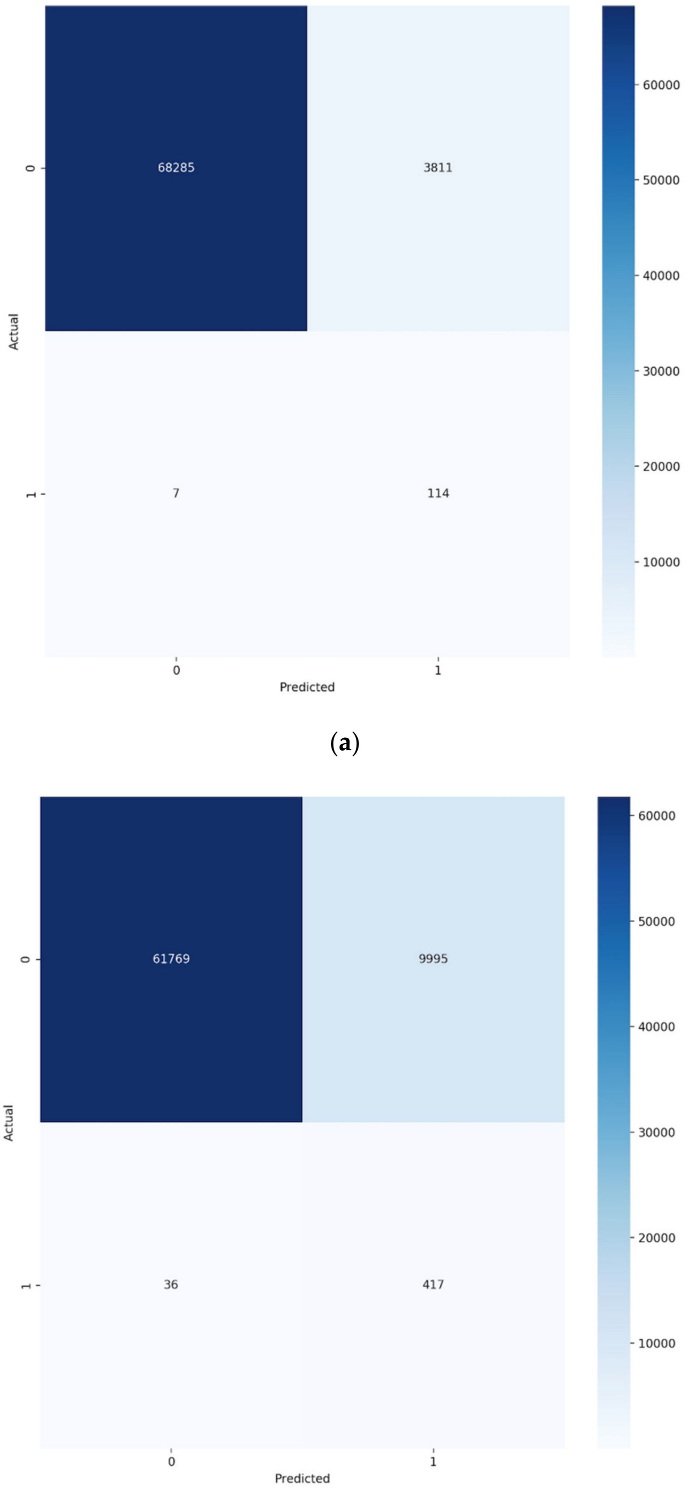
(**b**)

Figure 5. (**a**) Calibration plots for predicting in-hospital mortality within 48 h (**b**) Calibration plots for predicting overall in-hospital mortality.



(**a**) (**b**)

Figure 6. (**a**) Precision-Recall (PR) curve for in hospital mortality within 48 h (**b**) Precision-Recall (PR) curve for overall in-hospital mortality.



(**a**)

(**b**)

Figure 7. (**a**) Confusion matrix of the ensemble model 48-h in-hospital mortality predictions using Youden's index cut-off (**b**) Confusion matrix of the ensemble model predictions for overall in-hospital mortality using Youden's index cut-off.

Table 4. Metrics of the LR-tf-idf—gradient boosting ensemble model for predicting in-hospital mortality within 48 h.

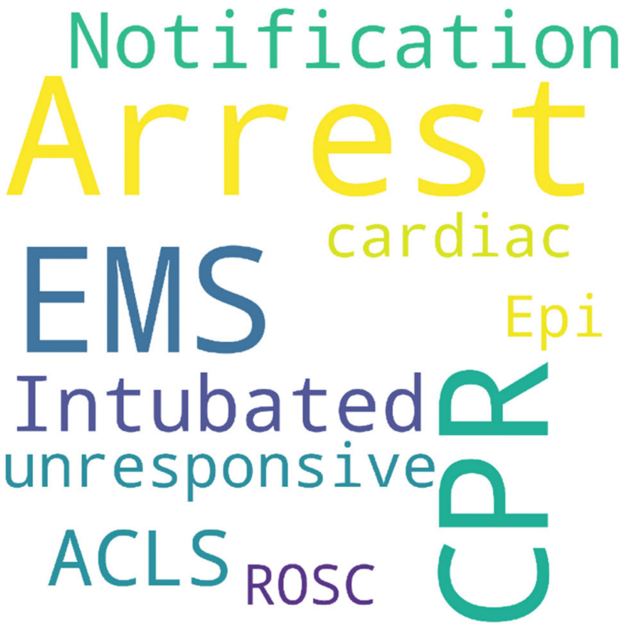
Fixed Specificity	Sensitivity	Specificity	Accuracy	FPR	PPV	NPV	F1
Youden's index	0.95 (95% CI: 0.91–0.98)	0.95	0.95 (95% CI: 0.95–0.95)	1:20	0.03 (95% CI: 0.02–0.03)	1.00 (95% CI: 1.00–1.00)	0.05 (95% CI: 0.04–0.06)
90%	0.96 (95% CI: 0.92–0.99)	0.90	0.89 (95% CI: 0.89–0.89)	1:10	0.02 (95% CI: 0.01–0.02)	1.00 (95% CI: 1.00–1.00)	0.03 (95% CI: 0.03–0.04)
99%	0.74 (95% CI: 0.66–0.82)	0.99	0.99 (95% CI: 0.99–0.99)	1:100	0.11 (95% CI: 0.09–0.13)	1.00 (95% CI: 1.00–1.00)	0.19 (95% CI: 0.15–0.22)

Abbreviations: FPR false positive rate; PPV positive predictive value; NPV negative predictive value; CI confidence interval.

Table 5. Metrics of the logistic regression-tf-idf–gradient boosting ensemble model for predicting overall in-hospital mortality.

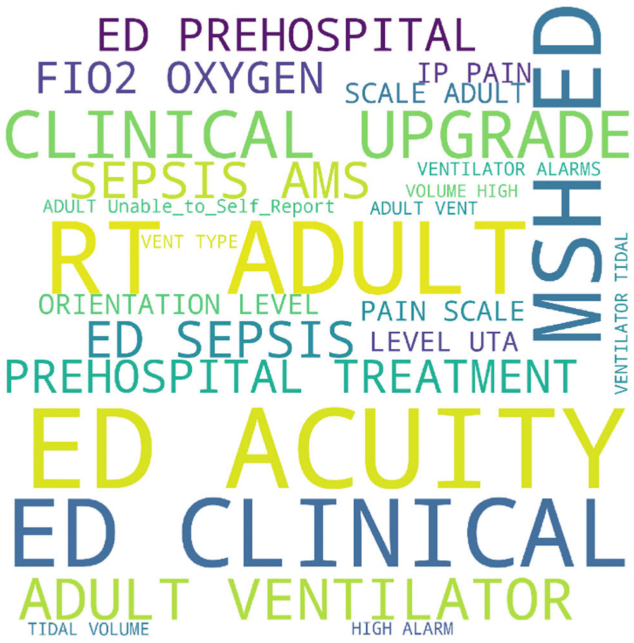
Fixed Specificity	Sensitivity	Specificity	Accuracy	FPR	PPV	NPV	F1
Youden's index	0.96 (95% CI: 0.94–0.98)	0.83	0.83 (95% CI: 0.83–0.84)	1:5.9	0.03 (95% CI: 0.03–0.04)	1.00 (95% CI: 1.00–1.00)	0.06 (95% CI: 0.06–0.07)
90%	0.87 (95% CI: 0.83–0.90)	0.90	0.89 (95% CI: 0.89–0.89)	1:10	0.05 (95% CI: 0.05–0.06)	1.00 (95% CI: 1.00–1.00)	0.10 (95% CI: 0.09–0.11)
99%	0.39 (95% CI: 0.35–0.43)	0.99	0.98 (95% CI: 0.98–0.99)	1:100	0.20 (95% CI: 0.17–0.22)	1.00 (95% CI: 1.00–1.00)	0.26 (95% CI: 0.23–0.29)

Abbreviations: FPR false positive rate; PPV positive predictive value; NPV negative predictive value; CI confidence interval.



(**a**)

Figure 8. *Cont.*



(**b**)

Figure 8. (**a**) Word cloud presentation of words related to 48-h in-hospital mortality. (**b**) Word Cloud of flowsheet terms related to 48-h in-hospital mortality.

4. Discussion

A simple method can be used for grasping complex Epic EHR data. Specifically, semi-tabular data can be represented as free-text, using the BoW approach. A model trained on this representation showed potential for predicting in-hospital mortality.

In the ED, many life and death decisions are performed under stressful conditions. Decision support tools can be helpful in this setting. Importantly, ED physicians have shown low accuracy (AUC 0.73) for predicting 30-day mortality [14].

Raita et al. trained several machine-learning models for predicting mortality and ICU admission at triage. They used data from 135k ED visits. Their variables were structured EHR features, including demographics, vital signs, chief complaints and comorbidities. The best model was a neural network with AUC 0.86 [9]. Klug et al. used gradient boosting to predict mortality at triage. They trained the model on 800k ED visits. Structured features included demographics, arrival mode, chief complaint, vital signs, ESI, previous visits and comorbidities. Their model showed AUC 0.96 for 48-h mortality [8].

In our study, an ensemble model showed an AUC 0.99 for 48-h mortality and AUC 0.96 for overall in-hospital mortality, albeit at the price of a low PPV. Predicting in-hospital mortality at triage is a "needle in a hay stack" problem. Typically, 2:1000 patients die within 48 h and 7:1000 die overall. For a sensitivity of 95%, the PPV was only 0.03, because of the low incidence of mortality. Yet, the FPR is 1:20. This means that in an ED with 200 patients per day, the algorithm will alert only 10 times. For an average physician who sees 20 patients per shift, the model will alert one patient per shift. This is important to prevent alert fatigue [15].

For FPR of 1:100, the sensitivity lowers to 74%, and the PPV rises to 0.11. With this threshold, the model will alert in the entire ED only twice a day. It will alert once per five physician shifts. Yet, the model will still detect approximately three quarters of early mortality cases.

We employed a common NLP technique to handle semi-structured data. Using BoW collections made it easy to represent the data as a sparse matrix. This is a memory efficient solution for utilizing Big-Data. The virtual/dense size of the dataset was 400 k visits × 120 k items per visit (including all possible unique free-text words and flowsheet term-value combinations). This amounts to 48 × 109 (50 billion) memory points. In the sparse matrix representation, each row physically occupied about 200 items. This amounts to 400 k × 200 = 80 × 106 (80 million) actual memory points. The sparse representation is 0.2% of the volume of the dense representation.

A previous study by Rajkomar et al. presented a method to extract the content of the EHR using the FHIR (Fast Healthcare Interoperability Resources) format [16]. FHIR was developed to represent clinical data in a consistent, hierarchical and extensible container format. Their study included a cohort of 216k hospitalized patients. The authors used a neural network to predict different outcomes [17]. In contrast with Rajkomar et al., we used a completely free-text-like representation of the data. This has the advantage of utilizing the entire dataset, as none of the data needs to be strictly encoded. The disadvantage of this approach is that it is site-specific.

We believe that the paradigm of "one solution to fit all sites" should be re-examined. At our site, there are almost 65k different unique flowsheet items with many values. The free-text records also have abbreviations and terms that are unique. We hypothesize that flexible solutions tailored to each medical institution are needed. This requires simple, efficient and flexible methods that can be easily implemented on-site. The presented method is an example of such a simple flexible solution. Future studies can elaborate on similar methods. Of note, the current method shows high predictive ability using logistic regression with tf-idf. This model is simple, fast, easy to implement and interpretable.

In a sense, converting data to "text," provides a very flexible way to "tell a story" about the data. For example, using this method, it was straightforward to create "past stories" for each patient. There are many possible ways to experiment with such stories, possibly improving results.

Limitations

This was a retrospective single-center study. A prospective applicative study is needed to prove the usefulness of the models. Yet, the essence of the study is the presentation of a simple method for harvesting data. Second, only in-hospital mortality was evaluated. Out of hospital death records were not available. Third, many different outcomes can be explored. For example, admittance to ICU or need for medical intervention. Fourth, we predicted all-cause mortality without stratifying to different pathologies. Fifth, there are multiple available machine and deep learning models. This study implemented two models. XGBoost, which shows state-of-the-art results in tabular tasks, and LR-tf-idf, which is a simple robust algorithm for NLP BoW. Other models can be explored. Sixth, a cut-off time of 30 min was used. We believe this is enough time for data to be accumulated, while still being clinically relevant as an early support tool. Any other cut-off time may be chosen. Seventh, a large number of variables (free-text words and flowsheet terms) were used. Even though data exploration such as we performed in Tables 3 and 4 can give insight, and although both logistic regression and gradient boosting are interpretable, the models may still be considered as black boxes.

5. Conclusions

A free-text-like approach can be useful for grasping large amounts of complex semi-structured EHR data.

Author Contributions: Conceptualization, E.K., M.A.L. and R.F.; data curation, E.K., M.A.L., A.Z., B.S.G., B.G.C., J.M., D.L.R. and R.F.; formal analysis, E.K., M.A.L., S.S., A.Z., B.S.G., B.G.C., J.M., D.L.R. and R.F.; methodology, E.K.; writing—original draft, E.K., B.S.G. and R.F.; writing—review and editing, M.A.L., S.S., A.Z., B.S.G., B.G.C., J.M., D.L.R. and R.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Anonymized participant data are held in a secure research server and will be handled in accordance with the ethical approval for this project.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Flowsheet terms associated with in-hospital mortality within 48 h and overall in-hospital mortality.

Term	Mutual Information	Mortality within 48 h (%)	Did not Die within 48 h (%)	Odds Ratio	*p* Value
In-hospital mortality within 48 h					
"ED ACUITY": 1	3.9	36.3	0.4	141.9	<0.001
"MSH ED CLINICAL UPGRADE": Yes	2.8	46.2	4	20.8	<0.001
"ED SEPSIS AMS": Yes	1.6	21.3	0.9	30.1	<0.001
"ED PREHOSPITAL TREATMENT": Yes	1.3	20.6	1.6	15.9	<0.001
"FIO2 OXYGEN": 100	1.2	10.1	0.1	140.6	<0.001
"IP PAIN SCALE ADULT": Unable_to_Self_Report	1.1	20.1	2	12.4	<0.001
"ORIENTATION LEVEL": UTA	1.1	13.9	0.5	31.6	<0.001
"RT (ADULT) VENT TYPE": Puritan_Bennett_PB840	1.1	9.8	0.1	118.9	<0.001
"RT (ADULT) VENTILATOR ALARMS ON": Y	1.1	9.5	0.1	115.4	<0.001
"RT (ADULT) VENTILATOR TIDAL VOLUME HIGH ALARM": 1000	1	9.4	0.1	120.9	<0.001
Overall in-hospital mortality					
"MSH ED CLINICAL UPGRADE": Yes	6.4	31.9	3.9	11.7	<0.001
"ED ACUITY": 2	5.3	57.4	22.2	4.7	<0.001
"ED ACUITY": 1	4.5	13.5	0.4	41.8	<0.001
"IP PAIN SCALE ADULT": Unable_to_Self_Report	3.3	16.2	1.9	9.9	<0.001
"ED SEPSIS AMS": Yes	3.3	12.8	0.8	17.2	<0.001
"MSH ED TEAM": Acute_One	2.7	33.3	14.5	2.9	<0.001
"MSH ED TEAM": Acute_Two	2.6	32.8	14.3	2.9	<0.001
"BP PATIENT POSITION": Supine	2.4	21	6.6	3.8	<0.001
"FALL RISK ACTION AT TRIAGE": Patient_placed_on_fall_precaution	2.3	19.4	5.7	4	<0.001
"ED SEPSIS INFECTION": Yes	2.3	19.3	5.7	4	<0.001

Table A2. Words associated with in-hospital mortality within 48 h and overall in-hospital mortality.

Word	Mutual Information	Mortality within 48 h (%)	Did not Die within 48 h (%)	Odds Ratio	*p* Value
In-hospital mortality within 48 h					
Arrest	3.5	28.9	0.2	259.1	<0.001
CPR	2.9	20.5	0	786.3	<0.001
EMS	2.7	58.2	8.9	14.2	<0.001
Notification	2.2	23.2	0.4	72	<0.001
Intubated	2.2	25.9	0.7	47.1	<0.001
ACLS	2.2	15.2	0	1072.3	<0.001
unresponsive	2.2	21.5	0.3	87.4	<0.001
cardiac	2.1	35	3	17.4	<0.001
Epi	2	18.3	0.2	114.3	<0.001
ROSC	1.8	12.4	0	856	<0.001
Overall in-hospital mortality					
EMS	4.6	34.8	8.8	5.5	<0.001
From	4.2	48.8	20.3	3.8	<0.001
Arrest	3.3	8.8	0.1	66.2	<0.001
Resus	3.1	15.8	2.1	8.8	<0.001
Per	3	30.9	11.1	3.6	<0.001
notification	2.8	9.5	0.4	26.5	<0.001
Admission	2.8	16.6	2.9	6.8	<0.001
CXR	2.7	20	4.8	4.9	<0.001
unresponsive	2.6	8.4	0.3	31.1	<0.001
CPR	2.6	5.9	0	221.6	<0.001

References

1. Obermeyer, Z.; Emanuel, E.J. Predicting the Future-Big Data, Machine Learning, and Clinical Medicine. *N. Engl. J. Med.* **2016**, *375*, 1216–1219. [CrossRef] [PubMed]
2. Carter, E.J.; Pouch, S.M.; Larson, E.L. The relationship between emergency department crowding and patient outcomes: A systematic review. *J. Nurs. Scholarsh. Off. Publ. Sigma Tau Int. Honor. Soc. Nurs.* **2014**, *46*, 106–115. [CrossRef] [PubMed]
3. Johnson, K.D.; Winkelman, C. The effect of emergency department crowding on patient outcomes: A literature review. *Adv. Emerg. Nurs. J.* **2011**, *33*, 39–54. [CrossRef] [PubMed]
4. Pines, J.M.; Iyer, S.; Disbot, M.; Hollander, J.E.; Shofer, F.S.; Datner, E.M. The effect of emergency department crowding on patient satisfaction for admitted patients. *Acad. Emerg. Med. Off. J. Soc. Acad. Emerg. Med.* **2008**, *15*, 825–831. [CrossRef] [PubMed]
5. Sun, B.C.; Hsia, R.Y.; Weiss, R.E.; Zingmond, D.; Liang, L.J.; Han, W.; McCreath, H.; Asch, S.M. Effect of emergency department crowding on outcomes of admitted patients. *Ann. Emerg. Med.* **2013**, *61*, 605–611. [CrossRef] [PubMed]
6. Chiu, I.M.; Lin, Y.R.; Syue, Y.J.; Kung, C.T.; Wu, K.H.; Li, C.J. The influence of crowding on clinical practice in the emergency department. *Am. J. Emerg. Med.* **2018**, *36*, 56–60. [CrossRef] [PubMed]
7. McHugh, M.; Tanabe, P.; McClelland, M.; Khare, R.K. More patients are triaged using the Emergency Severity Index than any other triage acuity system in the United States. *Acad. Emerg. Med. Off. J. Soc. Acad. Emerg. Med.* **2012**, *19*, 106–109. [CrossRef] [PubMed]
8. Klug, M.; Barash, Y.; Bechler, S.; Resheff, Y.S.; Tron, T.; Ironi, A.; Soffer, S.; Zimlichman, E.; Klang, E. A Gradient Boosting Machine Learning Model for Predicting Early Mortality in the Emergency Department Triage: Devising a Nine-Point Triage Score. *J. Gen. Intern. Med.* **2020**, *35*, 220–227. [CrossRef] [PubMed]
9. Raita, Y.; Goto, T.; Faridi, M.K.; Brown, D.F.M.; Camargo, C.A., Jr.; Hasegawa, K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit. Care* **2019**, *23*, 64. [CrossRef] [PubMed]
10. Available online: https://www.epic.com/about (accessed on 28 August 2021).

11. Barash, Y.; Guralnik, G.; Tau, N.; Soffer, S.; Levy, T.; Shimon, O.; Zimlichman, E.; Konen, E.; Klang, E. Comparison of deep learning models for natural language processing-based classification of non-English head CT reports. *Neuroradiology* **2020**, *62*, 1247–1256. [CrossRef] [PubMed]
12. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
13. Soffer, S.; Klang, E.; Barash, Y.; Grossman, E.; Zimlichman, E. Predicting in-hospital mortality at admission to the medical ward: A big-data machine learning model. *Am. J. Med.* **2021**, *134*, 227–234. [CrossRef] [PubMed]
14. Ouchi, K.; Strout, T.; Haydar, S.; Baker, O.; Wang, W.; Bernacki, R.; Sudore, R.; Schuur, J.D.; Schonberg, M.A.; Block, S.D.; et al. Association of Emergency Clinicians' Assessment of Mortality Risk with Actual 1-Month Mortality Among Older Adults Admitted to the Hospital. *JAMA Netw. Open* **2019**, *2*, e1911139. [CrossRef] [PubMed]
15. Co, Z.; Holmgren, A.J.; Classen, D.C.; Newmark, L.; Seger, D.L.; Danforth, M.; Bates, D.W. The tradeoffs between safety and alert fatigue: Data from a national evaluation of hospital medication-related clinical decision support. *J. Am. Med Inform. Assoc.* **2020**, *27*, 1252–1258. [CrossRef] [PubMed]
16. Tanaka, K.; Yamamoto, R. Implementation of a Secured Cross-Institutional Data Collection Infrastructure by Applying HL7 FHIR on an Existing Distributed EMR Storages. *Stud. Health Technol. Inform.* **2020**, *272*, 155–158. [CrossRef] [PubMed]
17. Rajkomar, A.; Oren, E.; Chen, K.; Dai, A.M.; Hajaj, N.; Hardt, M.; Liu, P.J.; Liu, X.; Marcus, J.; Sun, M.; et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* **2018**, *1*, 18. [CrossRef] [PubMed]