PLOS ONE

RESEARCH ARTICLE

Predicting Intensive Care Unit admission among patients presenting to the emergency department using machine learning and natural language processing

Marta Fernandes[1]*, Rúben Mendes[1], Susana M. Vieira[1], Francisca Leite[2], Carlos Palos[3], Alistair Johnson[4], Stan Finkelstein[5], Steven Horng[6], Leo Anthony Celi[4,7]

1 IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal, **2** Hospital da Luz Learning Health, Lisbon, Portugal, **3** Hospital Beatriz Ângelo, Luz Saúde, Lisbon, Portugal, **4** MIT Critical Data, Laboratory for Computational Physiology, Harvard-MIT Health Sciences & Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **5** Institute for Data, Systems and Society, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **6** Department of Emergency Medicine / Division of Clinical Informatics / Center for Healthcare Delivery Science, Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States of America, **7** Division of Pulmonary Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States of America

* marta.fernandes@tecnico.ulisboa.pt

OPEN ACCESS

Citation: Fernandes M, Mendes R, Vieira SM, Leite F, Palos C, Johnson A, et al. (2020) Predicting Intensive Care Unit admission among patients presenting to the emergency department using machine learning and natural language processing. PLoS ONE 15(3): e0229331. https://doi.org/10.1371/journal.pone.0229331

Editor: Ivan Olier, Liverpool John Moores University, UNITED KINGDOM

Received: August 17, 2019

Accepted: February 4, 2020

Published: March 3, 2020

Data Availability Statement: The data sets used in this study were de-identified and derived from the electronic medical records from patient visits to the Emergency Department at Beth Israel Deaconess Medical Center (BIDMC) and Hospital Beatriz Ângelo (HBA). Researchers must enter into a data sharing agreement with the covered entities (BIDMC and HBA) to obtain the data sets. Researchers can contact the BIDMC Committee on Clinical Investigation [1] and Grupo Luz Saúde [2] to execute a data sharing confidentiality agreement

Abstract

The risk stratification of patients in the emergency department begins at triage. It is vital to stratify patients early based on their severity, since undertriage can lead to increased morbidity, mortality and costs. Our aim was to present a new approach to assist healthcare professionals at triage in the stratification of patients and in identifying those with higher risk of ICU admission. Adult patients assigned Manchester Triage System (MTS) or Emergency Severity Index (ESI) 1 to 3 from a Portuguese and a United States Emergency Departments were analyzed. Variables routinely collected at triage were used and natural language processing was applied to the patient chief complaint. Stratified random sampling was applied to split the data in train (70%) and test (30%) sets and 10-fold cross validation was performed for model training. Logistic regression, random forests, and a random undersampling boosting algorithm were used. We compared the performance obtained with the reference model—using only triage priorities—with the models using additional variables. For both hospitals, a logistic regression model achieved higher overall performance, yielding areas under the receiver operating characteristic and precision-recall curves of 0.91 (95% CI 0.90-0.92) and 0.30 (95% CI 0.27-0.33) for the United States hospital and of 0.85 (95% CI 0.83-0.86) and 0.06 (95% CI 0.05-0.07) for the Portuguese hospital. Heart rate, pulse oximetry, respiratory rate and systolic blood pressure were the most important predictors of ICU admission. Compared to the reference models, the models using clinical variables and the chief complaint presented higher recall for patients assigned MTS/ESI 3 and can identify patients assigned MTS/ESI 3 who are at risk for ICU admission.

for the data set to be released. The BIDMC data is slated for public release in April 2020. Access to the HBA data must be requested from the hospital. [1] https://www.bidmc.org/research/research-and-academic-affairs/clinical-research-atbidmc/committee-on-clinical-investigation-irb/contact-us [2] https://www.hbeatrizangelo.pt/pt/institucional/informacao-de-privacidade/.

Funding: This work was supported by the Portuguese Foundation for Science & Technology (FCT) (URL 1), through IDMEC, under LAETA, project UIDB/50022/2020 and LISBOA-01-0145-FEDER-031474 supported by Programa Operacional Regional de Lisboa by FEDER (URL 2) and FCT. The work of Marta Fernandes was supported by the PhD Scholarship PD/BD/114150/2016 from FCT. URL 1: https://www.fct.pt/ URL 2: https://www.europarl.europa.eu/factsheets/pt/sheet/95/el-fondo-europeo-de-desarrollo-regional-feder- The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Emergency departments (EDs) often form the front line of health care systems and play a critical role in ensuring an efficient and quality service for patients with acute conditions [1]. The first evaluation, where the patient condition and acuity level are defined, is performed at triage, which has emerged as a method to identify patients who need immediate care.

According to a report published in 2017 by the Agency for Healthcare Research and Quality (AHRQ) [2], there were 120 million ED visits in 2006 in the United States and by 2014, there were 137.8 million ED visits, an increase of 14.8 percent. In 2015, the Organisation for Economic Co-operation and Development (OECD) Health Committee publish a report stating that the number of ED visits across OECD countries was about 31 per 100 population in 2011. The number of visits per capita was the highest in Portugal, with over 70 visits per 100 population [1]. Thus, the case study of Portugal is relevant, given that it represents high demand for emergency services, compared to other OECD countries.

The Manchester Triage System (MTS) and the Emergency Severity Index (ESI) are 5-level triage systems widely used in Europe and in the US, respectively. In ESI, treatment priority is decided on the basis of disease severity and the expected resource needs [3]. This system uses an algorithm, with ratings ranging from level 1 (patients with life-threatening conditions) to level 5 (the least resource-intensive patients). MTS priorities range from level 1 (emergent patients that should have immediate medical observation) to level 5 (non urgent patients that should wait a maximum time of 4 hours for medical observation).

Recent studies have shown good results in prediction of hospital admission [4–12], ED LOS [13], ICU admission [9, 14], mortality [9, 15, 16] and combined outcome of mortality and ICU admission [9, 14, 17] using machine learning techniques and historical information accessible from the EHR of triaged patients. There are also contributions of prediction models for triage classification in the literature [18–22]. Among the predictors used in the referred studies were age, gender, arrival mode, vital signs acquired at triage, chief complaint, time of admission, patient comorbidities and relevant medical history.

In this work, we employed machine learning to identify ED patients with high risk of ICU admission. We used data routinely collected at triage from the EDs of Hospital Beatriz Ângelo (HBA) in Portugal and of Beth Israel Deaconess Medical Center (BIDMC) in the United States. The primary outcome was admission to the ICU or equivalent in the first 24 hours after triage, accounting for differences in clinical practice across the two sites. At BIDMC, the outcome measure was admission to ICU. At HBA, the outcome measure consisted of admission to ICU or Intermediate Care Unit where medical or surgical patients who need ongoing monitoring are admitted. BIDMC does not have an intermediate care unit; all equivalent patients are admitted to the ICU. The models were developed for the cohort of patients assigned MTS/ESI 1 to 3. The patients assigned MTS/ESI 4 to 5 present to the ED for minor issues such as rashes or minor lacerations, and rarely are admitted to the ICU. We excluded these less urgent patients to reduce the class imbalance. We then compared the performance against a reference model trained only with the triage priority assigned to patients, with ESI or MTS, from the United States and Portuguese hospitals, respectively.

Materials and methods

Data acquisition

Data were acquired from the Emergency Department Information Systems (EDIS) of a Portuguese and a United States hospital. The data ranges from 2012 to 2016 and from 2011 to 2016, for the Portuguese and United States data, respectively, with a total of 599276 and 267257 ED

visits in the adult population ($\geq$ 18 years old). This study was approved by HBA Ethics Committee. The use of the BIDMC data was approved by the Institutional Review Board (IRB) under protocol number 2011P-000356. The HBA Ethics Committee and the BIDMC IRB waived the requirement for informed consent.

Predictors

For modelling, we included the variables routinely collected at triage (vital signs—temperature, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, mean arterial blood pressure, pulse oximetry SpO_2 and pain scale), the chief complaint, glycemia levels, Glasgow coma scale (GCS), the triage priority assigned to the patient, the patient age and gender, mode of arrival to the ED (ambulance, walk-in), disabilities (stretcher, wheelchair or none), time of triage (weekday, hour and month), ED visit (first triage registered on the system or not), prescription of complementary means of diagnostic at triage (number of exams), and type of exams prescribed (ophthalmology, otolaryngology, electrocardiogram, X-ray and orthopedic).

Outcomes

The outcomes considered as outputs for the models were the following, in a period within 24 hours after triage:

- **BIDMC**: ICU—admission to ICU
- **HBA**: ICU&INT—admission to ICU or to Intermediate Care Unit

Inclusion and exclusion criteria

At HBA, re-triage is performed when a patient is triaged again after the initial triage for new assessment of parameters, change of priority, activation of clinical pathways, or introduction/correction of other information in the registry. Re-triages were excluded in the dataset for modelling, given that not all patients are re-triaged. There were patients which were transferred from HBA ED to another hospital ED. There was no information of outcome for these patients, therefore patients referred to another ED within the 24 hours were excluded. Patients who died before ED admission were excluded as well. Finally, patients assigned MTS/ESI 4 to 5 were excluded, as well as patients assigned a white priority (in HBA this priority is assigned for patients with less urgency for care), since the study focused on patients assigned MTS/ESI 1 to 3. For detailed exclusion criteria refer to S1 Appendix in supplementary materials.

Modeling

Modeling design. We used stratified random sampling to split the dataset into train (70%) and test (30%) sets so that class labels were balanced in each dataset. The dataset was pre-processed before the modeling stage, which can be depicted in supplementary materials (S2 Appendix). We performed a stratified 10 fold cross validation (CV) in the training set to perform a randomized search for hyperparameter optimization. The information regarding hyperparameter tuning is depicted in S1 Table. The configuration of the model with highest AUROC was selected as well as the corresponding threshold. This model was evaluated in the held-out test dataset. We performed 100 iterations of bootstrapping random sampling in 95% confidence intervals (CI) to measure variance in performance. The methodology for modeling can be depicted in S1 Fig.

Modeling techniques. We used logistic regression (LR) with L2 regularization, RUSBoost and random forests regression bootstrap aggregation of decision trees (RFR). These algorithms

were selected so we could compare a boosting classification technique—RUSBoost, and a bagging regression technique—random forests with a more traditional technique—LR.

LR is a general statistical model originally developed by Joseph Berkson [23]. The prediction $\hat{y}$ or probability of an event for certain input features values x, is related to the N input features according to (1), with parameters $\beta_0, \beta_1, \ldots \beta_N$.

$$\hat{y}(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x + \ldots + \beta_N x)}} \tag{1}$$

RUSBoost is a well known boosting algorithm that uses a combination of RUS (random under-sampling) and the standard Adaptive Boosting (AdaBoost) procedure, and can improve the learning when using imbalanced data [24]. AdaBoost is adaptive in the sense that subsequent weak learners are adjusted in favor of those instances misclassified by previous classifiers. RUSBoost randomly removes examples from the majority class until the desired balance is achieved. This technique presents the advantage of being extremely fast to train the models, compared to other techniques such as LR and random forests, since the training dataset size is reduced. Denoting $\hat{y}$ as the boosting classifier prediction, with T as the total number of classifiers, each e_t is a weak learner that takes an object x as input and returns a value indicating its class. A set of weights ϖ is assigned for the T classifiers, in order to take a weighted average of their estimates. A learner with a good classification result will be assigned a higher weight than a poor one.

$$\hat{y}(x) = \sum_{t=1}^{T} \varpi_t e_t(x) \tag{2}$$

Random forests [25] perform a randomized sampling process to train a set of individual decision trees, aggregating the output to produce a single probabilistic prediction for each outcome. For classification tasks, the random forests classifier outputs the class which is voted more times by the individual trees. For regression tasks it gives the mean prediction of the individual trees as indicated in (3).

$$\hat{y}(x) = \frac{1}{T} \sum_{t=1}^{T} e_t(x) \tag{3}$$

We also applied a multimodeling approach with a majority voting classifier, where the voting was performed with all the predicted labels from the base learners, and the final prediction was made using the label with most votes. The decision criterion to select which models should vote for the final classification consisted on their individual performance, namely their sensitivity. Multimodeling approaches have been used in the literature [26–31] and a more comprehensive analysis of multimodeling and ensemble techniques can be found in [32].

Natural language processing. The chief complaint for BIDMC dataset is essentially semi-structured text mapped to SNOMED-CT using the HierArchical Presenting Problem ontologY (HaPPy) [33], which improves the quality of this feature. For this feature, contractions were fixed, punctuation was removed, words were set to lowercase and tokenized. We also performed abbreviation expansion, replaced numbers by words, removed stopwords and finally applied lemmatization. The chief complaint for HBA dataset consists of unstructured free written text and it was subjected to lowercasing, a process of temporal normalization, tokenization, abbreviations expansion and correction using Jaro-Winkler and stemming. More detailed preprocessing can be depicted in S2 Appendix.

To include the chief complaint as model predictor, the Term frequency–inverse document frequency (TF-idf) was used for text vectorization. Tf-idf is a numerical statistic that reflects how important a word is to a document in a collection or corpus and it was first introduced in [34]. The Tf–idf value increases proportionally to the number of times an N-gram (in the present case—a word) appears in a document and is offset by the number of documents in the corpus that contain the word. This process automatically adjusts the weighting of words that appear more frequently and which might have less meaning. The term frequency (tf) in Tf-idf expressed in (4) indicates how frequently a word appears in the document, measuring the local importance of it. The term inverse document frequency (idf) of each word is expressed in (5) and it measures the rareness of a term. Tf-idf is the product of tf and idf as expressed in (6).

$$tf(N-gram) = \frac{Number\ of\ times\ the\ N-gram\ appears\ in\ the\ Document}{Number\ of\ N-grams\ in\ the\ Document} \tag{4}$$

$$idf(N-gram) = \log_{10}\left(\frac{Number\ of\ documents}{Number\ of\ documents\ containing\ the\ N-gram}\right) \tag{5}$$

$$Tf-idf(N-gram) = tf(N-gram) \times idf(N-gram) \tag{6}$$

The number of N-grams (words) to select from each patient chief complaint as well as the total number of words to use from the training vocabulary are indicated in S1 Table.

Performance measures. According to [35], one of the most commonly reported measures for validating modeling performance, is the area under the receiver-operating characteristic curve (AUROC). This is a function of the true positive ratio or recall versus the false positive ratio (FPR), integrated over all thresholds. FPR in (8) corresponds to a false alarm ratio of the model and represents the cases where the patient is incorrectly classified as positive. Recall in (7) corresponds to the sensitivity of the model and represents the cases where the patient is correctly classified as being positive. An AUROC of 0.50 is achieved through random predictions where 1 represents a perfect discrimination. The pair ($Recall_k$, FPR_k) is referred to as an operating point for this curve.

The area under the precision-recall curve (AUPRC) was also assessed and it is a useful measure of success of prediction when the classes are very imbalanced. The AUPRC shows the trade-off between precision in (9) and recall in (7) for different thresholds. A high area under the curve represents both high recall and high precision. The pair ($Recall_k$, $Precision_k$) is referred to as an operating point for this curve. Other measures for assessing the modeling performance were the specificity or true negative rate (TNR) in (10), precision or positive predictive value (PPV) in (9) and accuracy in (11), which were used in previous studies [4–6]. We also assessed F1-score in (12), a measure that displays the trade-off between recall and precision, suited for dealing with imbalanced datasets [13]. We used Cohen's Kappa (κ) [36] to analyze inter-rater reliability, which represents the agreement between two variables [37]. Cohen's Kappa is presented in (13) where p_o is the empirical probability of agreement on the label assigned to a sample, and p_e is the expected agreement when both raters assign labels randomly. Finally, we present the standardized mortality ratio (SMR) which in this case represents the ratio between the observed number of positive outcomes (ICU admission) predicted by the

model and the number of positive outcomes which would be expected.

$$Recall = \frac{TP}{TP + FN}; \tag{7}$$

$$FPR = \frac{FP}{FP + TN}; \tag{8}$$

$$Precision = \frac{TP}{TP + FP}; \tag{9}$$

$$Specificity = \frac{TN}{TN + FP}; \tag{10}$$

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}; \tag{11}$$

$$F - score = \frac{2 \times TP}{2 \times TP + FN + FP}. \tag{12}$$

$$\kappa = \frac{p_0 - p_e}{1 - p_e}. \tag{13}$$

Where TN and TP indicate the true negatives and positives, patients that were correctly identified as belonging to class 0 and 1, respectively; FN and FP indicate the false negatives and positives, patients that were incorrectly identified as belonging to class 0 and 1, respectively.

Results

Emergency department data

In the BIDMC data with a population of 267257 adult ED visits, we excluded triages with unknown priority (n = 14341), obstetric patients (n = 5668), inconsistencies in vital signs (n = 25968), ESI-4 (n = 18188) and ESI-5 (n = 765) leaving a cohort of 120649 triaged patients, as presented in Fig 1. This cohort was comprised of 7.0% ESI-1, 34.2% ESI-2 and 58.8% ESI-3 patients. Among patients admitted to the ICU in the first 24 hours after triage (3426–2.8%), there were 35.9% ESI-1, 51.5% ESI-2 and 12.5% ESI-3.

In the HBA data with a population of 599276 adult ED visits, we excluded triaged patients with unknown age (n = 51), unknown priority (n = 473), unknown time of ED admission (n = 222), transfers to other hospital ED (n = 5524), obstetrics (n = 64130), MTS-4 (n = 287280), MTS-5 (n = 7100), white priority (n = 2095), activation of protocols (n = 20982), re-triages (n = 8515) and death before ED admission (n = 448) leaving a cohort of 235826 triaged patients, as presented in Fig 2. This cohort was comprised of 0.6% MTS-1, 17.5% MTS-2 and 81.9% MTS-3 patients. Among patients admitted to the ICU and Intermediate Care Unit in the first 24 hours after triage (1784–0.8%), there were 9.8% MTS-1, 53.3% MTS-2 and 37.0% MTS-3.

Demographics and a subset of variables are presented in Table 1. A list with all the variables used for modeling is presented in S2 Table. The descriptive statistics of all predictors used for modeling are presented in S3, S4 and S5 Tables. For both hospitals, the gender was balanced and triaged population had a median age of 59 and 51 years old in HBA and BIDMC datasets. In HBA dataset, the top five most common triage discriminators were moderate pain (35.0%),

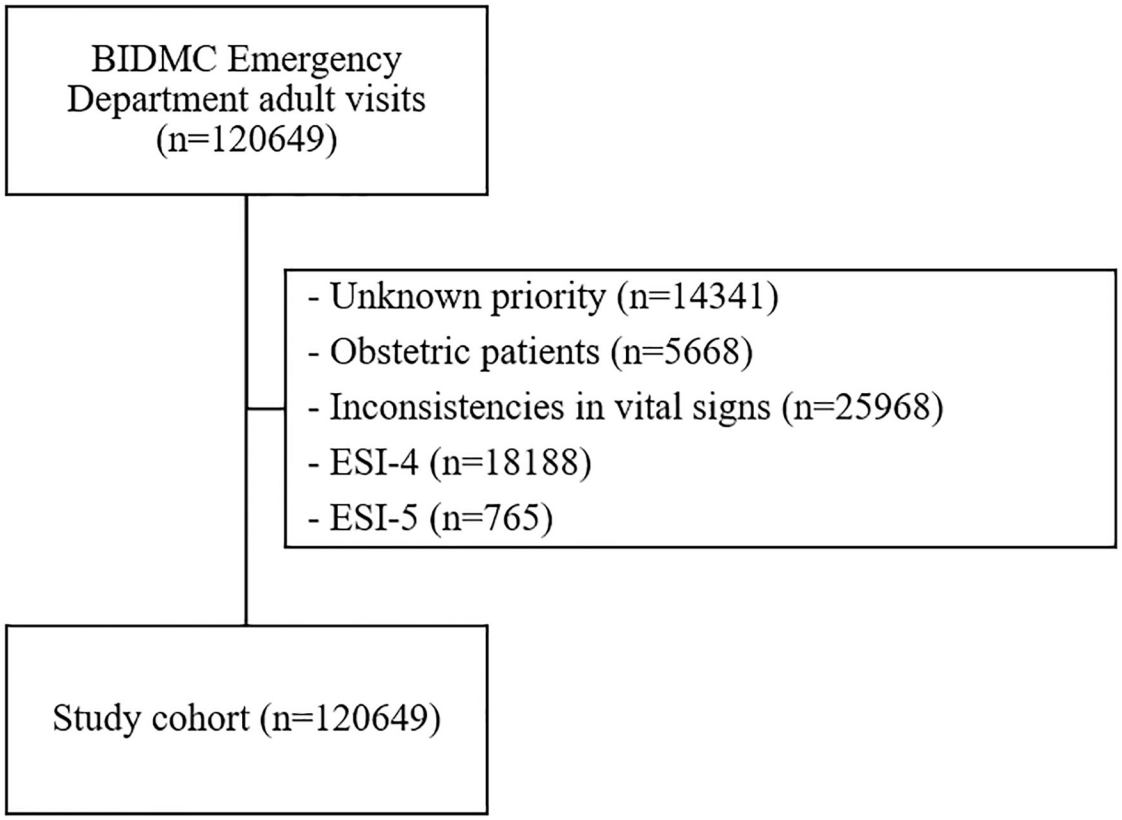

Fig 1. Inclusion and exclusion criteria for Beth Israel Deaconess Medical Center dataset. "n" corresponds to the number of triages.

https://doi.org/10.1371/journal.pone.0229331.g001

correspondent to a level of pain between 5 and 7; pleuritic pain (6.0%); sudden onset (5.2%) e.g. evidence of stroke; low pulse oximetry (4.7%) where $90\% \leq SpO_2 \leq 95\%$; and severe pain (4.6%) correspondent to a level of pain between 8 and 10. In BIDMC dataset, according to ICD-9 and ICD-10 codes assignment at triage, the top five most common conditions were chest pain (4.2%), abdominal pain (2.6%), syncope (2.0%), headache (2.0%) and pneumonia (1.4%). For both hospitals, the admitted patients were older with an average of 65 years old and there were more male patients being admitted than female. Compared to non-admitted, admitted patients to BIDMC and HBA ICU presented an average of: 2 and 1 breaths per minute higher; 9 and 5 beats per minute higher; 1% and 2% lower pulse oximetry; 7 and 3 mmHg lower mean arterial blood pressure. The average temperature of 37 degrees Celsius was the same for both admitted and not admitted patients in both hospitals.

Prediction of ICU admission in BIDMC

For prediction of ICU admission among BIDMC patients assigned the ESI 1 to 3, we had a training set with control/exposure groups of 82056/2398 patients, and the test set with 35167/1028 patients. The modeling results for ESI and additional features can be depicted in S7 Table and visualized graphically in Fig 3. Since the assignment of triage priority is subjective and can be variable across institutions, the model selected for further analysis was developed with all available predictors except for the triage priority. The additional features added to the BIDMC model were information on the number of abnormal and missing vital signs, mean arterial

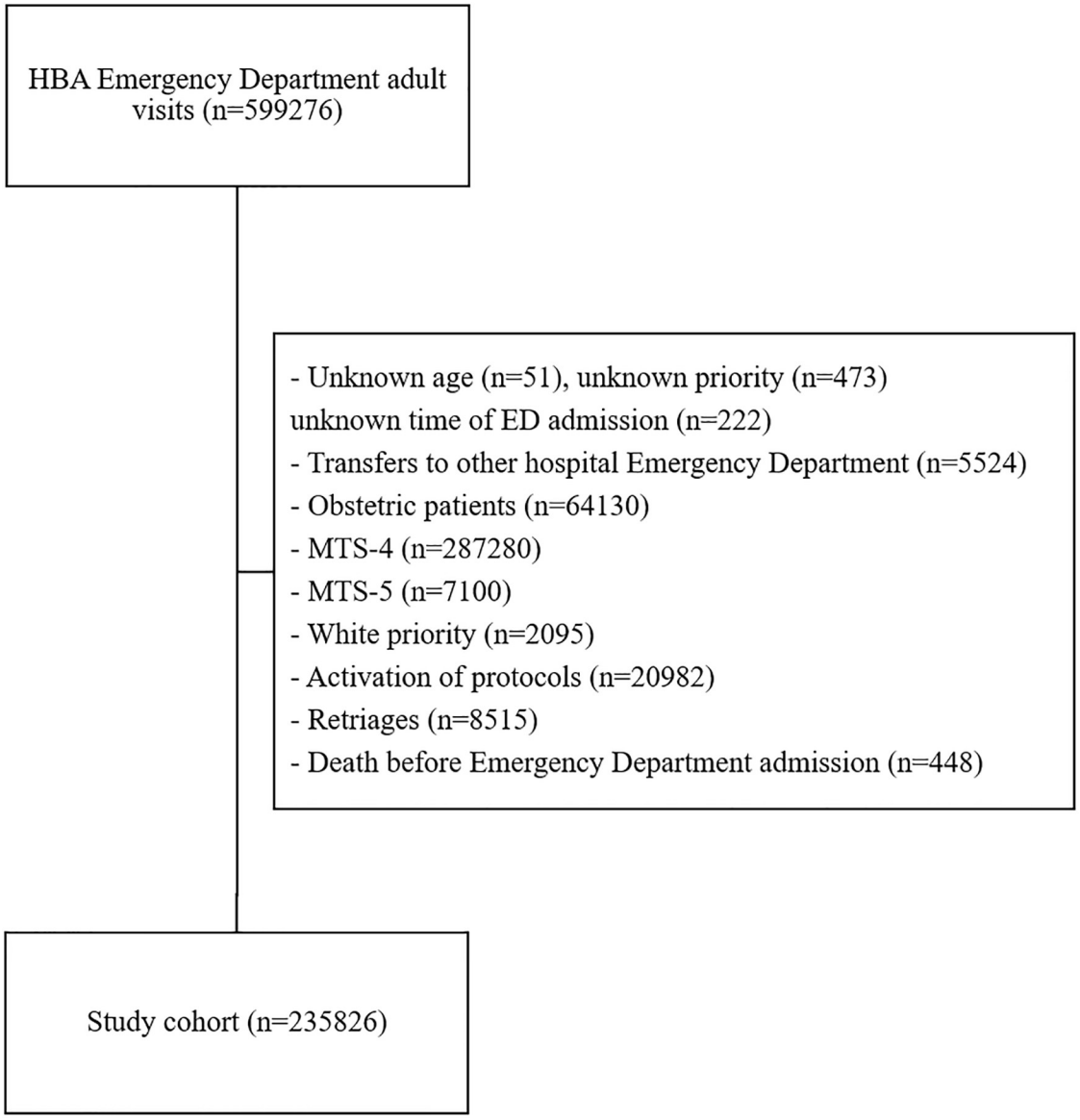

Fig 2. Inclusion and exclusion criteria for Hospital Beatriz Ângelo dataset. "n" corresponds to the number of triages.

https://doi.org/10.1371/journal.pone.0229331.g002

blood pressure, information regarding time of triage (weekday, hour and month) and the chief complaint.

Regarding the measure of F1-score and the AP, the values were relatively low due to the class imbalance present in the data (S7 Table). Analyzing the performance results obtained when adding clinical variables to ESI, the performance was overall higher when compared to the reference model using only ESI. Compared to RUSBoost or RFR, the LR model consistently presented higher sensitivity. RFR models had the tendency to overfit, while RUSBoost model presented higher F1-score results but at the cost of much lower sensitivity. When adding the chief complaint to the model using clinical variables, there was a significant increase in overall performance.

We assessed importance estimates of predictors through the absolute values of the coefficients from LR, as presented in Fig 4. The most important predictor was heart rate, followed

Table 1. Demographic variables and a subset of features available for both hospitals are summarized for each cohort of emergency department patients.

Variable (units)	BIDMC		HBA	
	ICU admission	No ICU admission	ICU admission	No ICU admission
Age (years old)	65 (19-93)	50 (19-93)	65 (18-101)	58 (18-108)
Female gender	1594 (47)	62502 (53)	757 (42)	130145 (56)
Male gender	1832 (53)	54721 (47)	1027 (58)	103903 (44)
Vital signs				
Respiratory rate (breaths/min)	19 (6-40)	17 (0-40)	18 (6-40)	17 (0-40)
Heart rate (beats/min)	93 (14-190)	84 (12-234)	91 (24-220)	86 (0-293)
Temperature (˚C)	37 (20-41)	37 (20-42)	37 (27-41)	37 (20-42)
Pulse oximetry (%)	97 (50-100)	98 (58-100)	94 (55-100)	96 (50-100)
Systolic blood pressure (mmHg)	126 (47-263)	135 (24-270)	138 (53-260)	143 (36-292)
Diastolic blood pressure (mmHg)	72 (16-214)	77 (6-191)	75 (25-140)	77 (6-201)
Mean Arterial Blood Pressure (mmHg)	90 (35-222)	97 (34-204)	96 (37-173)	99 (27-211)
Triage priorities				
Emergent (ESI-1, MTS-1)	1231 (36)	7173 (6)	174 (10)	1189 (1)
Very urgent (ESI-2, MTS-2)	1766 (52)	39540 (34)	950 (53)	40266 (17)
Urgent (ESI-3, MTS-3)	429 (12)	70510 (60)	660 (37)	192593 (82)
Outcome	3426 (2.8)	117223 (97.2)	1784 (0.8)	234048 (99.2)

The table shows number of patients. The figures in parentheses are the column percentages within each categorical variable for the respective outcome of admission. For continuous variables mean and range are presented. BIDMC—Beth Israel Deaconess Medical Center. HBA—Hospital Beatriz Ângelo. ESI—Emergency Severity Index. MTS—Manchester Triage System. ICU—Intensive Care Unit.

https://doi.org/10.1371/journal.pone.0229331.t001

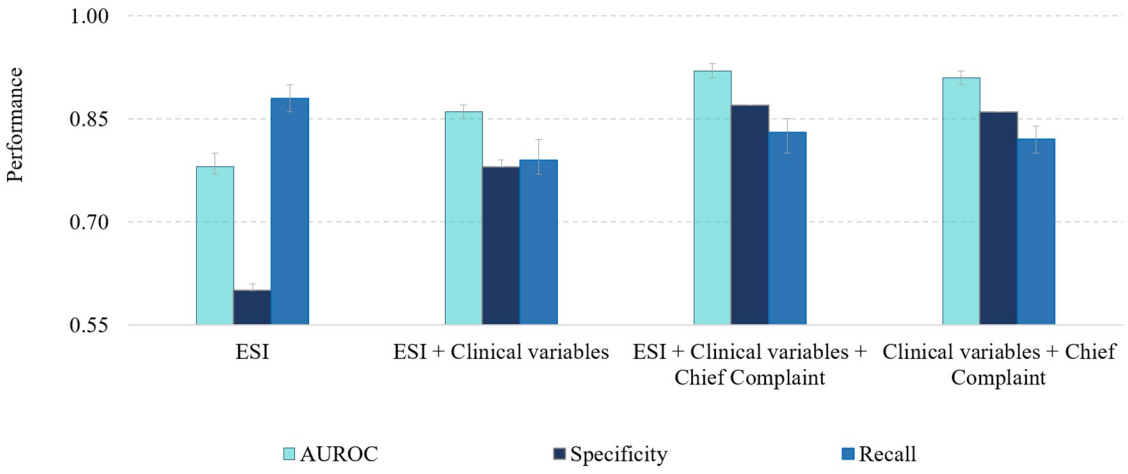

Fig 3. Performance of regularized logistic regression in test using the different subsets of predictors for Beth Israel Deaconess Medical Center dataset. ESI—Emergency Severity Index, AUROC—area under the ROC curve.

https://doi.org/10.1371/journal.pone.0229331.g003

by systolic blood pressure with a contribution in importance estimates of 90%, pulse oximetry with 80%, respiratory rate with 50% and the patient's age with 45%. The mean arterial blood pressure contributed an importance estimate of approximately 30% and the remaining predictors less than 20%. The time of triage, weekday and month contributed the least importance estimate of less than 1%.

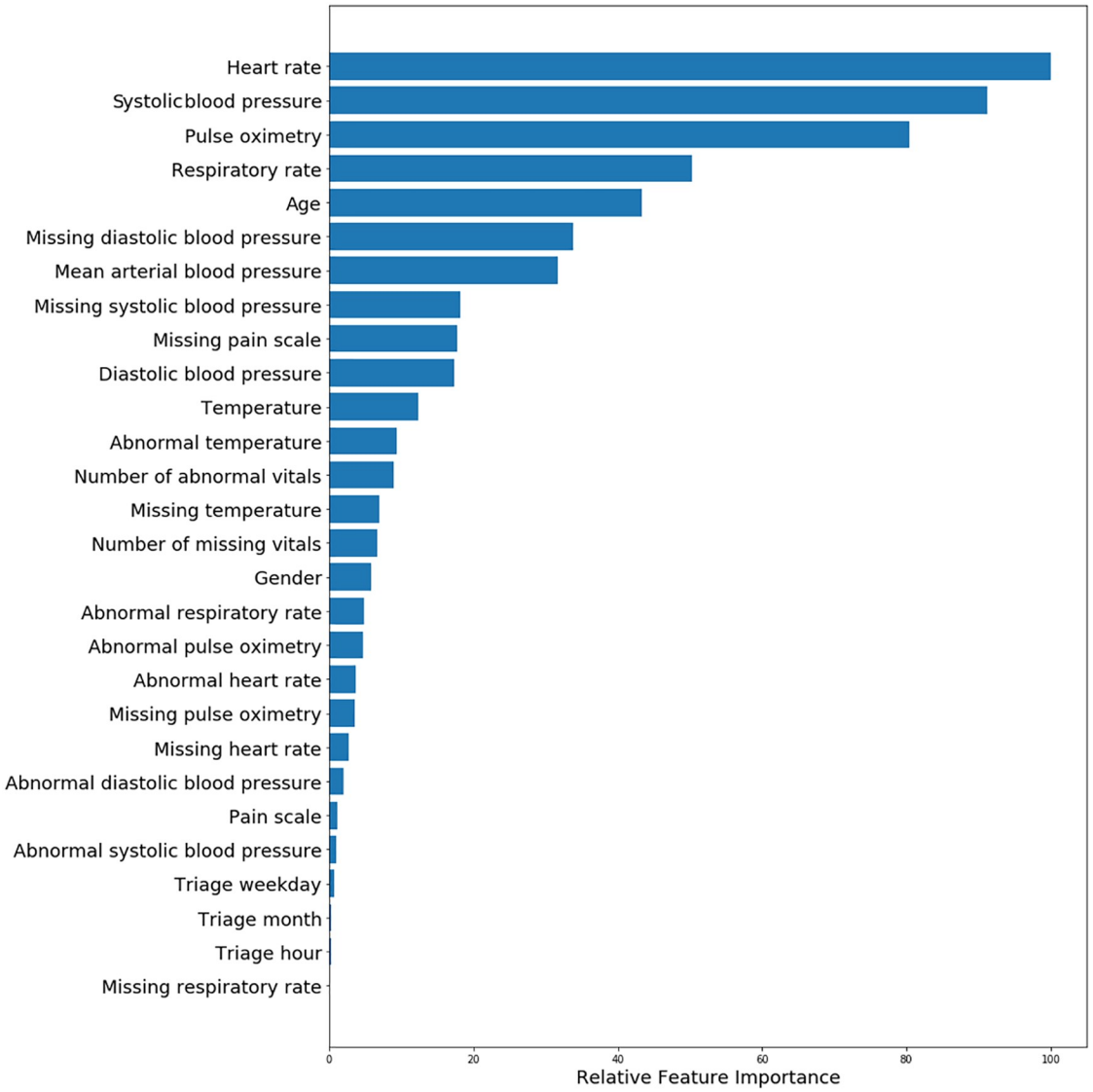

Fig 4. Relative importance of predictors of Intensive Care Unit admission for Beth Israel Deaconess Medical Center dataset obtained with regularized logistic regression using all available variables except triage priority.

https://doi.org/10.1371/journal.pone.0229331.g004

Prediction of ICU and Intermediate Care Unit admission in HBA

For prediction of ICU and Intermediate Care Unit admission among HBA patients assigned the MTS 1 to 3, we had a training set with control/exposure groups of 165082/1249 patients, and in the test set, 70750/535 patients. The modeling results for MTS and additional features can be depicted in S7 Table and visualized graphically in Fig 5. The F1-score and the precision values were low due to class imbalance. We observed that when adding the chief complaint, the sensitivity of the model decreased and the specificity increased. Therefore, a multi-model was created based on a voting classifier between the model using the chief complaint as predictor and the model not using this predictor. The multi-model could achieve a more balanced sensitivity and specificity. Since the assignment of triage priority is subjective, the multi-model consisted of the combination of models using all predictors except for the triage priority.

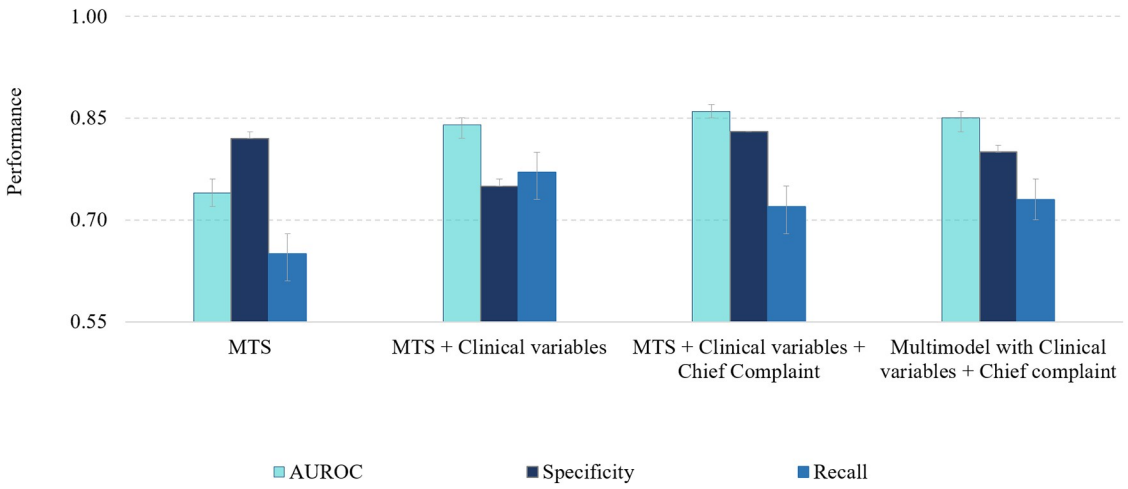

Fig 5. Performance of regularized logistic regression in test using the different subsets of predictors for Hospital Beatriz Ângelo dataset. MTS—Manchester Triage System, AUROC—area under the ROC curve.

https://doi.org/10.1371/journal.pone.0229331.g005

We assessed importance estimates of predictors through the absolute values of the coefficients given by LR, as presented in Fig 6. The most important predictor was the pulse oximetry, followed by glycaemia with a contribution in importance estimates of 90%, heart rate with 75%, orthopedic and ophthalmology consultations with 60% and 55%, respiratory rate with 50% and systolic blood pressure with 50%. The remaining predictors presented an importance estimate less than 30%. The time of triage, weekday and month contributed an importance estimate of less than 5%. These importance estimates are consistent with those of the BIDMC model, with the exception of age. The patient's age was ranked with a low importance of approximately 1% in the HBA model.

Models assessment and calibration

The models performance and calibration for both hospitals were assessed. Calibration curves, also referred to as reliability diagrams, present the fraction of patients in the positive class against the predicted probabilities. The mean of the predicted probabilities was computed for each decile. A well calibrated binary classifier will have an increasing number of true cases as one goes from the decile with the lowest mean predicted probability to the decile with the highest mean predicted probability. The performance and information regarding model parameters can be depicted in Table 2 where a comparison between the calibrated and non calibrated models is shown. The calibration curves for both models can be depicted in S2 Fig. Furthermore, we assessed the calibration curves for the different subsets of predictors for each hospital in S3 Fig.

For both hospitals, we observed an improvement in calibration as more features were included to the model. For HBA dataset, even in the best calibrated models, the risk of ICU admission was still over-estimated. This is acceptable for this specific use case, as it is better to over-estimate rather than under-estimate probabilities of ICU admission for the high-risk group of patients.

The AUROC and AUPRC for the BIDMC model with isotonic calibration and the HBA multi-model with no calibration are presented in Fig 7. For HBA, the multi-model without calibration was selected since it presented higher recall than the one with calibration.

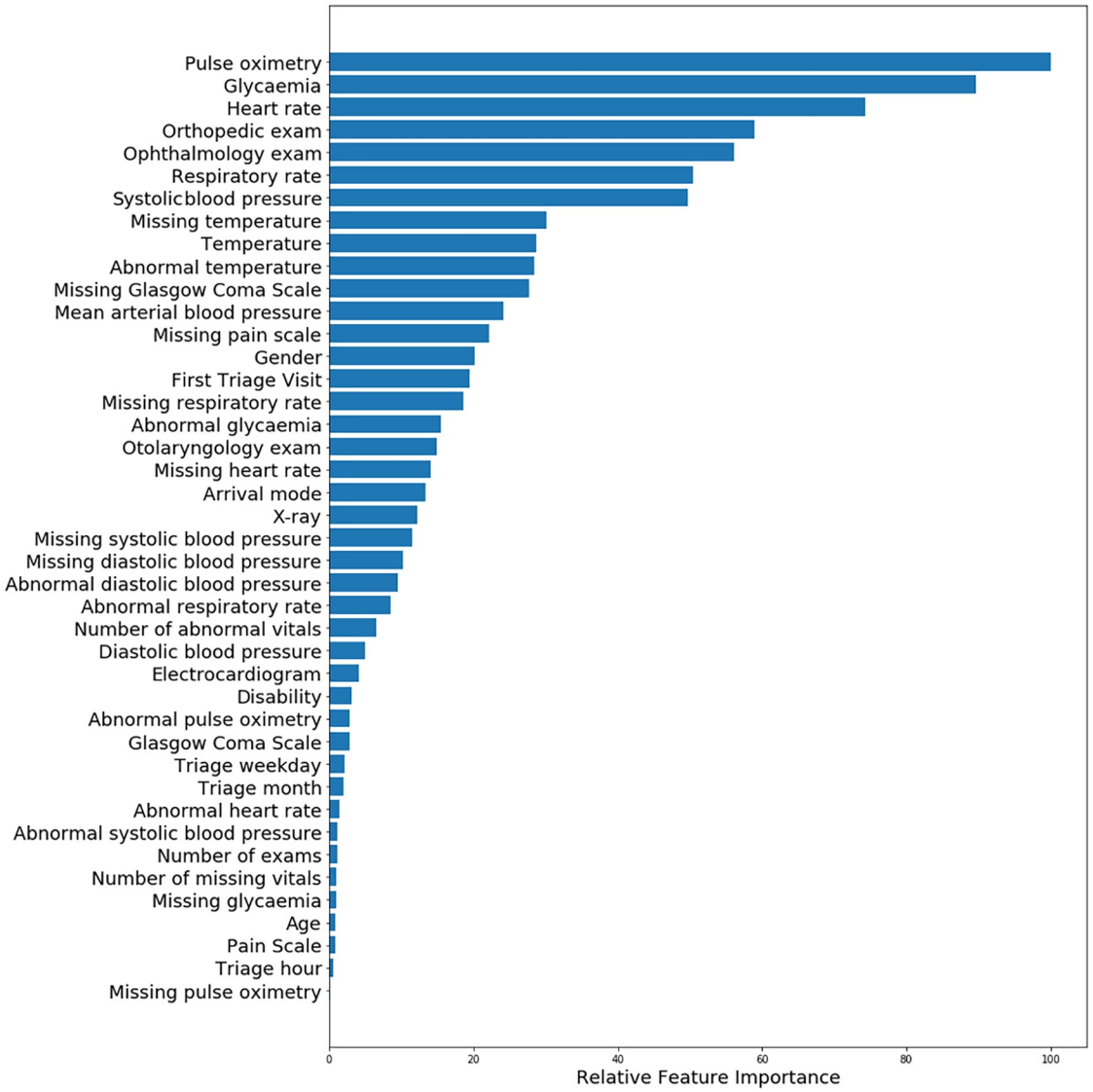

Fig 6. Relative importance of predictors of Intensive Care Unit and Intermediate Care Unit admission for Hospital Beatriz Ângelo dataset, obtained with regularized logistic regression using all available variables except triage priority. Exams are prescribed at the time of triage.

https://doi.org/10.1371/journal.pone.0229331.g006

Prediction models using clinical variables and the chief complaint were developed for both hospitals. For BIDMC, a model developed with clinical variables and the chief complaint presented lower recall in the identification of patients admitted to the ICU assigned ESI 1 and 2. However, the model was able to identify patients assigned an ESI-3 who were admitted to the ICU, contrary to the reference ESI (S4 Fig). For HBA, we developed a multi-model combining the model developed with clinical variables and the chief complaint with a model developed only with clinical variables so a balance between sensitivity and specificity could be attained. The scenario for HBA was similar as the one of BIDMC. For MTS 1 and 2 the multi-model presented lower recall in the identification of patients admitted to the ICU, however it had the ability to identify patients assigned an MTS-3 who were admitted to the ICU, contrary to the reference MTS (S5 Fig).

Table 2. Modeling results comparison between calibrated models.

	BIDMC model with no priority		HBA multi-model with no priority	
	No calibration	Isotonic calibration	No calibration	Isotonic calibration
AUROC	0.91 [0.90-0.92]	0.91 [0.90-0.92]	0.85 [0.83-0.86]	0.85 [0.83-0.86]
AUPRC	0.32 [0.28-0.35]	0.30 [0.27-0.33]	0.06 [0.05-0.07]	0.06 [0.05-0.07]
AP	0.31 [0.28-0.35]	0.30 [0.27-0.33]	0.06 [0.05-0.07]	0.06 [0.05-0.07]
Specificity	0.86 [0.86-0.87]	0.86 [0.86-0.86]	0.81 [0.80-0.81]	0.85 [0.84-0.85]
Recall	0.81 [0.79-0.84]	**0.82 [0.80-0.84]**	**0.73 [0.7-0.76]**	0.68 [0.65-0.72]
SMR	5.44 [5.1-5.78]	5.62 [5.28-5.97]	26.27 [24.27-28.5]	25.66 [23.68-27.84]
Threshold	0.035	0.030	0.009	0.007
N-gram range	Combination of unigrams and bigrams		M1: unigrams	
Words from vocabulary	9500		M1: 29000	
Warm start	No		Yes	
Regularization constant	1		1	

M1 corresponds to model which uses chief complaint for prediction. In brackets is the result for 100 bootstrapping iterations in 95% confidence intervals. The models selected have recall highlighted in bold. BIDMC—Beth Israel Deaconess Medical Center. HBA—Hospital Beatriz Ângelo. AUROC—area under the ROC curve. AUPRC—area under the precision recall curve. AP—Average precision. SMR—Standardized Mortality Ratio.

https://doi.org/10.1371/journal.pone.0229331.t002

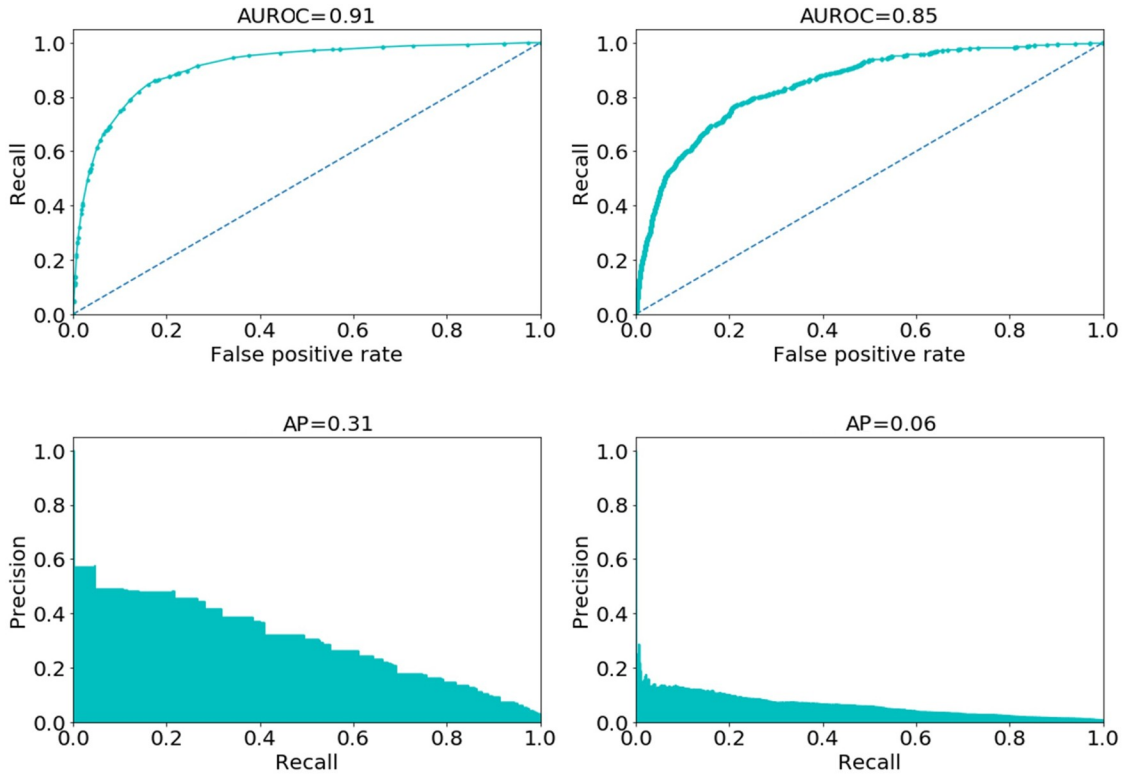

Fig 7. Performance of the model with isotonic calibration for Beth Israel Deaconess Medical Center (on the left) and of the multi-model with no calibration for Hospital Beatriz Ângelo (on the right) using in both logistic regression with all predictors except priority. AUROC—Area under the ROC curve. AUPRC—Area under the precision recall curve.

https://doi.org/10.1371/journal.pone.0229331.g007

Discussion

In this work, we developed models to predict risk of admission to the ICU at time of emergency department triage in two hospitals, one in Portugal and another in the US. Regularized logistic regression, random forests regression and random undersampling boosting of decision trees were used. The important predictors among patients assigned MTS/ESI 1 to 3 were identified. The performance of the models was compared to the reference model using only the ESI (for BIDMC) and MTS (for HBA) priority. The discrimination and calibration of the models were presented.

Since the assignment of triage priority is subjective and can be variable across institutions, the final models were developed using all clinical variables and the chief complaint, with the exception of triage priority. For both hospitals, heart rate, pulse oximetry, respiratory rate and systolic blood pressure ranked highly for prediction of ICU admission by the LR models. The model using semi-structured data from BIDMC for the chief complaint selected about a third of the words in the training vocabulary compared to the model using unstructured data from HBA. Although the free-text chief complaint may encode additional information, the amount of information compared to the structured chief complaint data could not be measured. For future work we propose to perform feature selection using the chief complaint and analyze the importance of this feature.

For BIDMC, a model developed with clinical variables and the chief complaint with isotonic calibration presented higher overall performance. For HBA, a multi-model combining the model developed with clinical variables and the chief complaint with a model developed only with clinical variables achieved a good balance between recall and specificity. The model for BIDMC data exhibited good calibration properties while for the HBA data, the multi-model over-estimated this risk. We concluded that for this case, an over-estimation of the probabilities of ICU admission for the high-risk group of patients can be better than an under-estimation of the risk. The models presented a higher recall in the identification of patients admitted to the ICU in MTS/ESI-3 priority level while the reference MTS/ESI presented higher recall for MTS/ESI 1 and 2 priority levels. The low measures of F1-score and precision were due to the class imbalance present in the data.

In a similar study [9] several machine learning models were developed to predict a critical outcome of admission to ICU or in-hospital death and their performance was compared with that of the ESI. The machine learning models outperformed the ESI reference model (e.g., AUROC, 0.86 (95%CI 0.85–0.87) in a deep neural network vs 0.74 (95%CI 0.72–0.75) in the reference model). The most important predictors of the critical outcome were patient's age, respiratory rate, heart rate, systolic blood pressure, pulse oximetry and arrival by ambulance. When compared to the BIDMC model, the same variables were identified as the most important predictors. In another study [14], a LR model was developed to predict a critical outcome of admission to ICU or in-hospital death in a cohort of patients aged 75 and older. The most important predictors were respiratory rate, systolic blood pressure, pulse oximetry and the Glasgow Coma Score. Except for the Glasgow Coma Score—which was not available in the BIDMC dataset, the same vital signs were ranked as the most important predictors.

In a third paper [8], a triage tool called "e-triage" was developed using random forests to predict the need for critical care, an emergency procedure, and inpatient hospitalization. The e-triage models had an AUROC ranging from 0.73 to 0.92 and demonstrated equivalent or improved prediction of patient outcomes compared with ESI at different EDs. Similar to this study, the models developed for BIDMC and HBA were able to predict ICU admission at the time of ED triage among patients assigned MTS/ESI-3 priority level. These results demonstrate an opportunity to complement the already existing triage systems with machine learning models and avoid under-triaging.

Supporting information

S1 Appendix. Exclusion criteria detailed.
(PDF)

S2 Appendix. Data pre-processing.
(PDF)

S1 Table. Hyperparameter optimization in random search cross validation.
(PDF)

S2 Table. Variables used for modelling Hospital Beatriz Ângelo (HBA) and Beth Israel Deaconess Medical Center (BIDMC) emergency department data. (*) Additional predictors available only for HBA dataset.
(PDF)

S3 Table. Demographics and vital signs variables used for modelling Hospital Beatriz Ângelo and Beth Israel Deaconess Medical Center emergency department data. The table shows number of patients. The figures in parentheses are the column percentages within each categorical variable for the respective outcome of admission. For continuous variables mean and range are presented.
(PDF)

S4 Table. Statistics of the additional Hospital Beatriz Ângelo predictor variables used for modelling. The table shows number of patients. The figures in parentheses are the column percentages within each categorical variable for the respective outcome of admission.
(PDF)

S5 Table. Additional variables used for modelling both hospitals emergency departments data. The table shows number of patients. The figures in parentheses are the column percentages within each categorical variable for the respective outcome of admission.
(PDF)

S6 Table. Criteria for outlier exclusion and abnormal values identification. DBP—diastolic blood pressure. (1) values not within normal range.
(PDF)

S7 Table. Average modeling performance results in test. In brackets is the result for 100 bootstrapping iterations in 95% confidence intervals.
(PDF)

S1 Fig. Methodology steps for modeling. AUROC—area under the ROC curve, AUPRC—area under the precision recall curve, TNR—true negative rate or specificity. RUSBoost—Random undersampling boosting algorithm. Tf-idf—Term frequency–inverse document frequency.
(TIF)

S2 Fig. Calibration curves of the logistic regression model with isotonic calibration for Beth Israel Deaconess Medical Center (on top) and of the multi-model for Hospital Beatriz Ângelo (on bottom), using all available predictors. Annotations with labels are presented only for the selected models.
(TIF)

S3 Fig. Calibration curves for Beth Israel Deaconess Medical Center (on the right) and for Hospital Beatriz Ângelo (on the left) according to the different subsets of modeling predictors.
(TIF)

S4 Fig. Confusion matrix for the models using data from Beth Israel Deaconess Medical Center.
(TIF)

S5 Fig. Confusion matrix for the models using data from Hospital Beatriz Ângelo.
(TIF)

S6 Fig.
(TIF)

Acknowledgments

The authors would like to acknowledge both hospitals Beth Israel Deaconess Medical Center and Hospital Beatriz Ângelo for having provided access to their databases for this study. There are no conflicts of interest.

Author Contributions

Conceptualization: Marta Fernandes, Susana M. Vieira, Francisca Leite, Carlos Palos, Stan Finkelstein, Steven Horng, Leo Anthony Celi.

Data curation: Marta Fernandes, Rúben Mendes, Alistair Johnson, Steven Horng.

Formal analysis: Marta Fernandes.

Funding acquisition: Marta Fernandes, Susana M. Vieira, Stan Finkelstein.

Investigation: Marta Fernandes, Rúben Mendes, Susana M. Vieira, Alistair Johnson, Steven Horng.

Methodology: Marta Fernandes, Rúben Mendes, Susana M. Vieira, Steven Horng, Leo Anthony Celi.

Software: Marta Fernandes, Rúben Mendes, Alistair Johnson, Steven Horng.

Validation: Marta Fernandes, Alistair Johnson, Steven Horng, Leo Anthony Celi.

Visualization: Marta Fernandes.

Writing – original draft: Marta Fernandes.

Writing – review & editing: Marta Fernandes, Susana M. Vieira, Alistair Johnson, Steven Horng, Leo Anthony Celi.

References

1. Berchet C, et al. Emergency Care Services: Trends, Drivers and Interventions to Manage the Demand. OECD Publishing; 2015.
2. Moore B, Stocks C and Owens P. Trends in Emergency Department Visits, 2006–2014, Hcup Statistical Brief# 227. Rockville, MD: Agency for Healthcare Research and Quality. Retrieved from https://www.hcup-us.ahrq.gov/reports/statbriefs/sb227-Emergency-Department-Visit-Trends.pdf.
3. Dong SL, Bullard MJ, Meurer DP, Blitz S, Holroyd BR and Rowe BH. The effect of training on nurse agreement using an electronic triage system. Canadian Journal of Emergency Medicine. 2007; 9 (4):260–266. https://doi.org/10.1017/s1481803500015141 PMID: 17626690
4. Araz Ozgur M., Olson David and Ramirez-Nafarrate Adrian. Predictive analytics for hospital admissions from the emergency department using triage information. International Journal of Production Economics. 2019; 208: 199–207. https://doi.org/10.1016/j.ijpe.2018.11.024
5. Graham B, Bond R, Quinn M and Mulvenna M. Using data mining to predict hospital admissions from the emergency department. IEEE Access. 2018; 6:10458–10469. https://doi.org/10.1109/ACCESS.2018.2808843

6. Parker CA, Liu N, Wu SX, Shen Y, Lam SSW and Ong MEH. Predicting hospital admission at the emergency department triage: A novel prediction model. The American Journal of Emergency Medicine. 2019; 37(8):1498–1504. https://doi.org/10.1016/j.ajem.2018.10.060 PMID: 30413365

7. Hong WS, Haimovich AD and Taylor RA. Predicting hospital admission at emergency department triage using machine learning. PloS One. 2018; 13(7):e0201016. https://doi.org/10.1371/journal.pone.0201016 PMID: 30028888

8. Levin S, Toerper M, Hamrock E, Hinson JS, Barnes S, Gardner H, et al. Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index. Annals of Emergency Medicine. 2018; 71(5):565–574. https://doi.org/10.1016/j.annemergmed.2017.08.005 PMID: 28888332

9. Raita Y, Goto T, Faridi MK, Brown DF, Camargo CA and Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Critical Care. 2019; 23(1):64. https://doi.org/10.1186/s13054-019-2351-7 PMID: 30795786

10. Zlotnik A, Alfaro MC, Pérez MCP, Gallardo-Antolín A and Martínez JMM. Building a Decision Support System for Inpatient Admission Prediction With the Manchester Triage System and Administrative Check-in Variables. CIN: Computers, Informatics, Nursing. 2016; 34(5):224–230. https://doi.org/10.1097/CIN.0000000000000230 PMID: 26974710

11. Cameron A, Rodgers K, Ireland A, Jamdar R and McKay GA. A simple tool to predict admission at the time of triage. Emergency Medicine Journal. 2015; 32(3):174–179. https://doi.org/10.1136/emermed-2013-203200 PMID: 24421344

12. Sun Y, Heng BH, Tay SY and Seow E. Predicting hospital admissions at emergency department triage using routine administrative data. Academic Emergency Medicine. 2011; 18(8):844–850. https://doi.org/10.1111/j.1553-2712.2011.01125.x PMID: 21843220

13. Azari A, Janeja VP and Levin S. (2015, November). Imbalanced learning to predict long stay Emergency Department patients. In 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 807-814). IEEE.

14. Barfod C, Lauritzen MMP, Danker JK, Sölétormos G, Forberg JL, Berlac PA, et al. Abnormal vital signs are strong predictors for intensive care unit admission and in-hospital mortality in adults triaged in the emergency department-a prospective cohort study. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine. 2012; 20(1):28. https://doi.org/10.1186/1757-7241-20-28 PMID: 22490208

15. Teubner DJ, Considine J, Hakendorf P, Kim S and Bersten AD. Model to predict inpatient mortality from information gathered at presentation to an emergency department: The Triage Information Mortality Model (TIMM). Emergency Medicine Australasia. 2015; 27(4):300–306. https://doi.org/10.1111/1742-6723.12425 PMID: 26147765

16. Coslovsky M, Takala J, Exadaktylos AK, Martinolli L and Merz TM. A clinical prediction model to identify patients at high risk of death in the emergency department. Intensive Care Medicine. 2015; 41(6):1029–1036. https://doi.org/10.1007/s00134-015-3737-x PMID: 25792208

17. LaMantia MA, Stewart PW, Platts-Mills TF, Biese KJ, Forbach C, Zamora E, et al. Predictive value of initial triage vital signs for critically ill older adults. Western Journal of Emergency Medicine. 2013; 14(5):453. https://doi.org/10.5811/westjem.2013.5.13411 PMID: 24106542

18. Azeez D, Ali MAM, Gan KB and Saiboon I. Comparison of adaptive neuro-fuzzy inference system and artificial neutral networks model to categorize patients in the emergency department. SpringerPlus. 2013; 2(1):416. https://doi.org/10.1186/2193-1801-2-416 PMID: 24052927

19. Azeez D, Gan K, Ali M and Ismail M. Secondary triage classification using an ensemble random forest technique. Technology and Health Care. 2015; 23(4):419–428. https://doi.org/10.3233/THC-150907 PMID: 25791174

20. Wang ST. Construct an optimal triage prediction model: A case study of the emergency department of a teaching hospital in Taiwan. Journal of Medical Systems. 2013; 37(5):9968. https://doi.org/10.1007/s10916-013-9968-x PMID: 23990379

21. Zmiri D, Shahar Y and Taieb-Maimon M. Classification of patients by severity grades during triage in the emergency department using data mining methods. Journal of Evaluation in Clinical Practice. 2012; 18(2):378–388. https://doi.org/10.1111/j.1365-2753.2010.01592.x PMID: 21166962

22. Aziz D, Ali MM, Gan K and Saiboon I. (2012, June). Initialization of adaptive neuro-fuzzy inference system using fuzzy clustering in predicting primary triage category. In 2012 4th International Conference on Intelligent and Advanced Systems (ICIAS2012) (Vol. 1, pp. 170-174). IEEE.

23. Cramer, Jan Salomon. The origins of logistic regression. (2002): 4.

24. Seiffert C, Khoshgoftaar TM, Van Hulse J and Napolitano A. RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 2010; 40(1):185–197. https://doi.org/10.1109/TSMCA.2009.2029559

25. Breiman L. Random forests. Machine Learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

26. Fernandes MP, Silva CF, Vieira SM and Sousa JM. (2014, July). Multimodeling for the prediction of patient readmissions in intensive care units. In 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1837-1842). IEEE.

27. Salgado CM, Azevedo CS, Garibaldi J and Vieira SM. (2015, August). Ensemble fuzzy classifiers design using weighted aggregation criteria. In 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-5). IEEE.

28. Erraguntla M, Zapletal J and Lawley M. Framework for Infectious Disease Analysis: A comprehensive and integrative multi-modeling approach to disease prediction and management. Health Informatics Journal. 2019; 25(4), 1170–1187. https://doi.org/10.1177/1460458217747112 PMID: 29278956

29. Salgado C, Fernandes M, Horta A, Xavier M, Sousa J and Vieira S. (2017, July). Multistage modeling for the classification of numerical and categorical datasets. In 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-6). IEEE.

30. Horta AB, Salgado C, Fernandes M, Vieira S, Sousa JM, Papoila AL, et al. Clinical decision support tool for Co-management signalling. International Journal of Medical Informatics. 2018; 113:56–62. https://doi.org/10.1016/j.ijmedinf.2018.02.014 PMID: 29602434

31. Xiao Y, Wu J, Lin Z and Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. Computer Methods and Programs in Biomedicine. 2018; 153:1–9. https://doi.org/10.1016/j.cmpb.2017.09.005 PMID: 29157442

32. Ju C, Bibaut A and van der Laan M. The relative performance of ensemble methods with deep convolutional neural networks for image classification. Journal of Applied Statistics. 2018; 45(15):2800–2818. https://doi.org/10.1080/02664763.2018.1441383 PMID: 31631918

33. Horng Steven and Greenbaum Nathaniel R and Nathanson Larry A and McClay James C and Goss Foster R and Nielson Jeffrey A. Consensus Development of a Modern Ontology of Emergency Department Presenting Problems—The Hierarchical Presenting Problem Ontology (HaPPy). Applied Clinical Informatics. 2019; 10(03):409–420. https://doi.org/10.1055/s-0039-1691842 PMID: 31189204

34. Luhn HP. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development. 1957; 1(4):309–317. https://doi.org/10.1147/rd.14.0309

35. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass). 2010; 21(1):128. https://doi.org/10.1097/EDE.0b013e3181c30fb2

36. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20(1):37–46. https://doi.org/10.1177/001316446002000104

37. McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica. 2012; 22 (3):276–282. https://doi.org/10.11613/BM.2012.031 PMID: 23092060