A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis

*William P.T.M. van Doorn[1,2], Patricia M. Stassen[3,4], Hella F. Borggreve[3], Maaike J. Schalkwijk[3], Judith Stoffers[3], Otto Bekers[1,2], Steven J.R. Meex[*1,2]*

Affiliations

[1] Department of Clinical Chemistry, Central Diagnostic Laboratory, Maastricht University Medical Center, Maastricht, The Netherlands.

[2] CARIM School for Cardiovascular Diseases, Maastricht University, Maastricht, The Netherlands.

[3] Department of Internal Medicine, Division of General Internal Medicine, Section Acute Medicine, Maastricht University Medical Centre, Maastricht University, Maastricht, The Netherlands.

[4] CAPHRI School for Care and Public Health Research Institute , Maastricht University, Maastricht, The Netherlands.

[*]Address for correspondence

Steven J.R. Meex, PhD

Central Diagnostic Laboratory

Maastricht University Medical Center+

PO Box 5800

6202 AZ Maastricht

The Netherlands

steven.meex@mumc.nl

Abstract: 289

Word count: 3527

Table and Figures: 5

References: 41

Abstract

Introduction: Patients with sepsis who present to an emergency department (ED) have highly variable underlying disease severity, and can be categorized from low to high risk. Development of a risk stratification tool for these patients is important for appropriate triage and early treatment. The aim of this study was to develop machine learning models predicting 31-day mortality in patients presenting to the ED with sepsis and to compare these to internal medicine physicians and clinical risk scores.

Methods: A single-center, retrospective cohort study was conducted amongst 1,344 emergency department patients fulfilling sepsis criteria. Laboratory and clinical data that was available in the first two hours of presentation from these patients were randomly partitioned into a development (n=1,244) and validation dataset (n=100). Machine learning models were trained and evaluated on the development dataset and compared to internal medicine physicians and risk scores in the independent validation dataset. The primary outcome was 31-day mortality.

Results: A number of 1,344 patients were included of whom 174 (13.0%) died. Machine learning models trained with laboratory or a combination of laboratory + clinical data achieved an area-under-the ROC curve of 0.82 (95% CI: 0.80-0.84) and 0.84 (95% CI: 0.81-0.87) for predicting 31-day mortality, respectively. In the validation set, models outperformed internal medicine physicians and clinical risk scores in sensitivity (92% vs. 72% vs. 78%;p<0.001,all comparisons) while retaining comparable specificity (78% vs. 74% vs. 72%;p>0.02). The model had higher diagnostic accuracy with an area-under-the-ROC curve of 0.85 (95%CI: 0.78-0.92) compared to abbMEDS (0.63,0.54-0.73), mREMS (0.63,0.54-0.72) and internal medicine physicians (0.74,0.65-0.82).

Conclusion: Machine learning models outperformed internal medicine physicians and clinical risk scores in predicting 31-day mortality. These models are a promising tool to aid in risk stratification of patients presenting to the ED with sepsis.

Introduction

Among emergency department (ED) presentations, a substantial number of patients present with symptoms of sepsis [1]. Sepsis is defined as a systemic inflammatory response syndrome (SIRS) to an infection and is associated with a wide variety of risks including septic shock and death [2]. Mortality rates of sepsis are as high as 16%, potentially increasing up to 40% when suffering from septic shock [2, 3]. Novel clinical decision support (CDS) systems capable of identifying low- or high-risk patients could become important for early treatment and triage of ED patients, but also for preventing unnecessary referrals to the intensive care unit (ICU). EDs are one of the most overcrowded units of a modern hospital, highlighting the importance of proper allocation and management of resources [1]. Development of a risk stratification tool for patients with sepsis may improve health outcome in this group, but may also contribute to resolve the problem of overcrowded EDs.

Currently, a wide variety of clinical risk scores are used in routine clinical care to facilitate risk stratification of patients with sepsis [4]. These include the relatively simple (quick) sequential organ failure assessment ((q)SOFA) score [5, 6], but also more complex scores such as the abbreviated Mortality in Emergency Department Sepsis (abbMEDS) score and modified Rapid Emergency Medicine Score (mREMS) [7, 8]. These traditional risk scores have shown varying performance for predicting 28-day mortality (area under the receiver operating characteristic curve (AUC) for abbMEDS: 0.62-0.85,mREMS: 0.62-0.84 and SOFA: 0.61-0.82) [3, 8-11]. In addition, clinical judgment of the attending physician in the ED plays an important role in risk stratification. The judgment of physicians was found to be a moderate to good predictor (AUC of 0.68-0.81) of mortality in the ED [12, 13].

Interestingly, a new group of CDS systems are being developed based on machine learning (ML) technology [14]. Machine learning can extract information from complex, non-linear data and provide insights to support clinical decision making. Hence, the first studies emerged that report machine learning-based mortality prediction models using data from patients with sepsis presenting to the ED [15-26]. Unfortunately, these studies did not provide a comparison with physicians in terms of prognostic performance. Recently, a new group of machine learning algorithms

termed gradient boosting trees emerged; showing superior performance compared to other ML models in some problems within the medical domain [27, 28]. Exploring if these models can outperform clinical risk scores and clinical judgment of physicians in their ability to identify low- or high-risk patients is a necessary step to explore the potential value of machine learning models in clinical practice.

The aim of this study was to develop machine learning-based prediction models for all-cause mortality at 31 days based on available laboratory and clinical data from patients presenting to the ED with sepsis. Subsequently, we compared the performance of these machine learning models with judgment of internal medicine physicians and clinical risk scores; abbMEDS, mREMS and SOFA.

Methods

Study design and setting

We performed a retrospective cohort study among all patients who presented to the ED at the Maastricht University Medical Centre+ between January 1, 2015 and December 31, 2016. All patients aged ≥18 years being referred to the internal medicine physician with sepsis, defined as a proven or suspected infection, and two or more SIRS and/or qSOFA criteria (S1 supporting information) were included in this study [2, 5, 29]. Patients with missing clinical data or with less than four laboratory results were excluded. Also, patients who refused to give consent were excluded. This study was approved by the medical ethical committee (METC 2019-1044) and the hospital board of the Maastricht University Medical Centre+. Furthermore, the study follows the STROBE guidelines and was conducted according to the principles of the Declaration of Helsinki [30]. The ethics committee waived the requirement for informed consent.

Data collection and processing

We collected clinical and laboratory data from all patients included in the study available within two hours after initial ED presentation. Clinical data were manually extracted through the electronic health record of the patient and included characteristics such as vital signs, hemodynamic parameters, and medical history (S1 Table). Biomarkers requested for standard clinical care were acquired through the laboratory information system. Biomarkers that were ordered in less than 1/1000 patients were excluded from the analysis. A list of included biomarkers is provided in S1 Table. Missing values did not require any processing as our machine learning model is capable of dealing with missing data. Instead, we created an additional variable for each biomarker with a discrete 'absence' or 'presence' feature to enable our model to distinguish between the absence and presence of a laboratory test within a patient. These features were included in both datasets. Finally, we derived two datasets from the processed data:

1. Laboratory dataset: this dataset consisted of age, sex, time of laboratory request and all requested laboratory biomarkers within two hours after the initial laboratory request

2. Laboratory + clinical dataset: this dataset contained all variables from the laboratory dataset, and additionally clinical, vital and physical (e.g. length and weight) characteristics of the patient

A full overview of all variables present in each dataset is described in S1 Table. Datasets were anonymized and randomly divided into two subsets: 1) a development subset (n=1,244), used for model training and evaluation, and 2) an independent validation subset (n=100), used for final validation and comparison of models with judgment of acute internal medicine physicians and clinical risk scores; abbMEDS and mREMS. A schematic overview of the study design and model development is depicted in Fig 1. Data processing and manipulation was performed using Python programming language (version 3.7.1) using packages numpy (version 1.17) and Pandas (version 0.24).

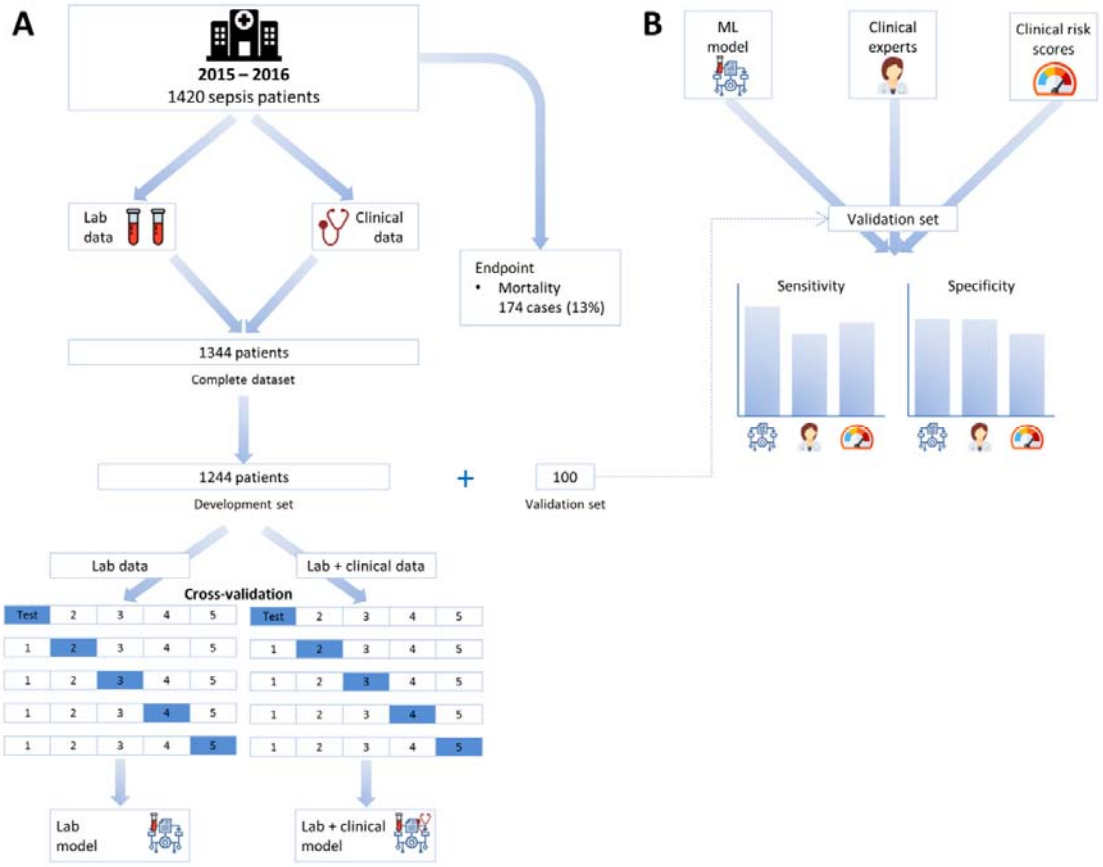

Fig 1. Overview of study design and model development. (A) We included 1,344 patients with a diagnosis of sepsis who presented to the ED. Patients were randomly partitioned in a development subset (n=1,244), used to train and evaluate performance of machine learning models, and a validation subset (n=100), used to compare models with internal medicine physicians and clinical risk scores. Cross-validation was used to obtain a robust estimate of model performance in the development subset. (B) The machine learning model with the highest cross-validation performance was compared internal medicine physicians and clinical risk scores to predict 31-days mortality.

Outcome measure

Septic shock during presentation was defined as systolic blood pressure (SBP) ≤90 mmHg and mean arterial pressure (MAP) ≤65 mmHg despite adequate fluid resuscitation. The outcome measure for this study was death within 31 days (1 month) after initial ED presentation. All-cause mortality information was acquired through electronic health records.

Model training and evaluation

Our proposed predictive model uses individual patient data available within two hours after initial ED presentation and generates the probability of mortality within 31 days. This prediction task can be solved by a variety of statistical and machine learning models. In the current study we evaluated logistic regression, random forest, multi-layer perceptron neural networks and XGBoost (S2 supporting information and S2 Table) on the laboratory dataset. We selected XGBoost as our machine learning model of choice as this was proven to possess the highest baseline performance (S2 Table). XGBoost is a recent implementation of gradient tree boosting systems which involve combining the predictions of many "weak" decision trees into a strong predictor [27]. This recent implementation is characterized by integral support of missing data and regularization mechanisms to prevent overfitting [27]. XGBoost models and their development can be altered by adjusting the parameters of the technique, referred to as "hyperparameters". Due to sample size limitations and the scope of our study, we decided not to optimize our hyperparameters and predefined them as described in S3 Table.

We employed stratified K-fold cross validation to assess the generalizability of our prediction models. Briefly, we randomly partitioned the development subset (n=1,244) into five, equally sized, folds. During each round of cross-validation, four of these folds were used to train our models ("train set") and the fifth was used to evaluate performance ("test set"). This was done in such a manner that every fold would be labeled as test set only once. We monitored training and test set errors to ensure that training increased performance on the test set. Accordingly, training was terminated after 5,000 rounds or when performance on the test set did not further improve for 10 rounds. We evaluated developed models trained with (i) the laboratory

dataset or (ii) the laboratory + clinical dataset, resulting in a total of two independent cross-validations.

Model explanation

To explain the output of our XGBoost models, we used the SHapley Additive exPlanations (SHAP) algorithm, to help us understand how a single feature affects the output of the model [31-33]. SHAP uses a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions [34, 35]. A Shapley value states, given the current set of variables, how much a variable in the context of its interaction with other variables contributes to the difference between the actual prediction and the mean prediction. That is, the mean prediction plus the sum of the Shapley values for all variables equals the actual prediction. It is important to understand that this is fundamentally different to direct variable effects known from e.g. (generalized) linear models. The SHAP value for a variable should not be seen as its direct -and isolated effect- but as its aggregated effect when interacting with other variables in the model. In our specific case, positive Shapley values contribute towards a positive prediction (death), whilst low or negative Shapely values contribute towards a negative prediction (survival). ML training and evaluation was done in Python using packages Keras (version 2.2.2), XGBoost (version 0.90), SHAP (version 0.34.0) and scikit-learn (version 0.22.1). The analysis code for this study is available on reasonable request.

Comparison of machine learning with internal medicine physicians and clinical risk scores

Performance of machine learning models was compared with clinical judgment of acute internal medicine physicians (n=4) and clinical risk scores in a validation subset of patients with sepsis (n=100) which were not previously exposed to the ML model. We selected the best performing machine learning model from cross-validation and trained this with identical hyperparameters as previously described on the full development subset. A machine learning prediction of higher than 0.50 was considered as a positive prediction. Next, we calculated the mREMS, abbMEDS and SOFA clinical risk scores as described previously (S1 supporting information) [8, 36].

Acute internal medicine physicians (n=4; 2 experienced consultants in acute internal medicine and 2 experienced residents acute internal medicine) were asked to predict 31-day mortality in the validation subset, based on retrospectively collected clinical and laboratory data. This data was presented in the form of a simulated electronic health record.

Statistical analysis

Descriptive analysis of baseline characteristics was performed using IBM SPSS Statistics for Windows (version 24.0). Continuous variables were reported as means with standard deviation (SD) or medians with interquartile ranges (IQRs) depending on the distribution of the data. Categorical variables were reported as proportions. Cross-validated models were assessed by receiver operating characteristic (ROC) curves and compared by their AUC using the Wilcoxon matched-pairs signed rank test. Besides diagnostic performance, we assessed calibration in cross-validations with reliability curves [37] and brier scores [38]. In our final validation subset, we compared the predictive performance of our best performing ML model to the judgment of acute internal medicine physicians and clinical risk scores with respect to sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy and AUC. Differences in AUC were tested using the method of DeLong et al [39]. Confidence intervals for proportions (e.g. sensitivity) were calculated using binomial testing and compared using McNemar's test. To analyze individual differences between internal medicine physicians, we performed two additional sensitivity analyses. First, the Cohen κ statistic was used to measure the inter-observer agreement between the internal medicine physicians. The level of agreement was interpreted as nil if κ was 0 to 0.20; minimal, 0.21 to 0.39; weak, 0.40 to 0.59; moderate, 0.60 to 0.79; strong, 0.80 to 0.90; and almost perfect, 0.90 to 1 [40]. Second, we compared the machine learning model against alternating groups of internal medicine physicians in which one physician was removed in each comparison.

Results

Study population and characteristics

During the study period, 5,967 patients presented to the ED who were referred to an internal medicine physician in our hospital. Of these patients, we included 1,420 patients with a suspected or proven infection, fulfilling the SIRS and/or qSOFA criteria. A number of 76 patients were excluded due to missing clinical data (n=23) and insufficient number of laboratory results (n=53), to form a final cohort of 1,344 patients (S1 Fig). Among all patients, 102 (7.6%) suffered from septic shock during presentation at ED and 174 (13.0%) died within 31 days after initial ED presentation. Baseline characteristics of the study patients in development and validation datasets are shown in Table 1.

Table 1. Baseline characteristics of patients in the development and validation datasets.

Characteristics	**Development** N = 1.244	**Validation** N = 100
Demographics		
Age	71.3 (58.8-82.3)	70.8 (58.4-82.8)
Sex, female	567 (45.6)	58 (58.0)
Comorbidity		
Cancer	446 (35.9)	28 (28.0)
Cardiopulmonary	381 (30.6)	30 (30.0)
Diabetes	264 (21.2)	19 (19.0)
Renal disease	128 (10.3)	9 (9.0)
Liver disease	42 (3.4)	7 (7.0)
Neuropsychiatric	65 (5.2)	2 (2.0)
Focus of infection at ED		
Respiratory tract	421 (33.8)	34 (34.0)
Urinary tract	218 (17.5)	18 (18.0)
Gastrointestinal tract	415 (33.4)	37 (37.0)
Others	75 (6.0)	6 (6.0)
Skin	115 (9.2)	5 (5.0)
Severity scores		
abbMEDS[a]	5.5 (3-8)	6 (3-8)
mREMS[b]	7 (6-9)	7 (6-9)
SOFA[c]	7 (5-9)	6 (5-8)
Outcomes		
Septic shock	94 (7.6)	8 (8.0)
31-day mortality	161 (12.9)	13 (13.0)

[a] AbbMEDS, Abbreviated Mortality in ED Sepsis, was calculated as described by Vorwerk et al [8].

[b] mREMS, modified Rapid Emergency Medicine Score, was calculated as described by Chang et al [36].

[c] SOFA, Sepsis-related Organ Failure Assessment, was calculated as described by Vincent et al [6].

Machine learning development and evaluation

To assess the generalizability of our developed XGBoost models, we employed five-fold cross validation on the development dataset (n=1,244). XGBoost models trained with laboratory data achieved an AUC of 0.82 (95% CI: 0.80 – 0.84) for predicting 31-day mortality (Fig 2). The performance improved, although not statistically significant, when clinical data was added to the laboratory data to train XGBoost to an AUC of 0.84 (95% CI: 0.81 – 0.87) for mortality (compared to lab only; p=0.25). Individual cross-validation results of each model are depicted in S2 Fig. Additionally, calibration curves show well calibrated models with brier scores between 0.08 to 0.10 (S3 Fig).

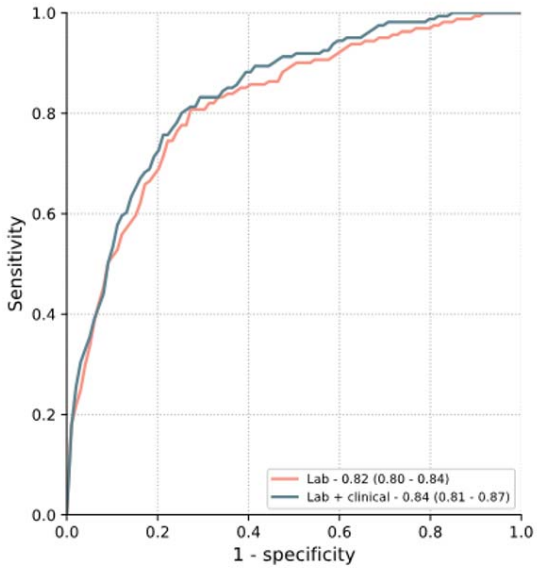

Fig 2. XGBoost model performance for predicting all-cause mortality at 31 days in the development dataset. Models trained with laboratory data achieved a mean AUC of 0.82 (95% CI: 0.80 – 0.84) for predicting 31-day mortality. Predictive performance increased when models were trained with laboratory + clinical data to a mean AUC of 0.84 (95% CI: 0.81 – 0.87), but this was not statistically different (p=0.25).

Model explanation

To identify which laboratory and clinical features contributed most to the performance of our models, we calculated SHAP values for the (i) laboratory and (ii) laboratory + clinical models (Fig 3). Among the highest ranked features, we observe features that are also often used in risk scores including urea, platelet count, glasgow coma score (GCS) and blood pressure. Interestingly, we also observe features such as glucose, lipase, and GCS which are less commonly associated with mortality in sepsis patients. An extended analysis of the correlation between important features in our models and risk scores is provided in S4 Table. Moreover, these SHAP plots allow us to examine the individual impact of laboratory and clinical features on the predictions of our models. For example, higher urea and C-reactive protein (CRP) levels (represented by red points) have a high SHAP value and thus a positive effect on the model outcome (death).

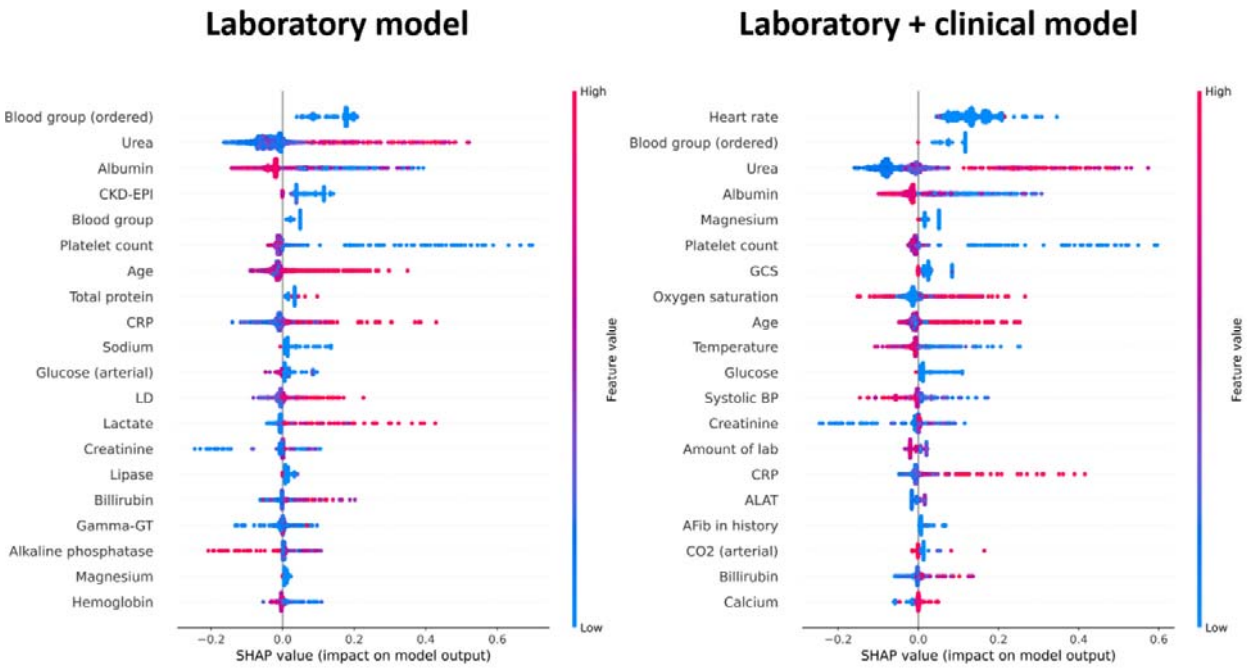

Fig 3. Analysis of parameter importance in the XGBoost models. Models with laboratory data (left) and with laboratory + clinical data (right) were analyzed using SHAP values. Individual parameters are ranked by importance in descending order based on the sum of the SHAP values over all the samples. Negative or low SHAP values contribute towards a negative model outcome (survival), whereas high SHAP values contribute towards a positive model outcome (death).

Machine learning versus internal medicine physicians and clinical risk scores

To explore the potential value of machine learning models in clinical practice, we compared the model trained with laboratory + clinical data with acute internal medicine physicians and clinical risk scores, abbMEDS, mREMS and SOFA, to predict 31-day mortality. In an independent validation subset (n=100) -which the model never had been exposed to before- it achieved a sensitivity of 0.92 (95% CI: 0.87-0.95, Fig 4A) and specificity of 0.78 (95% CI: 0.70-0.86, Fig 4B). In terms of sensitivity, the machine learning model significantly outperformed internal medicine physicians (0.72, 95% CI: 0.62-0.81; p<0.001), abbMEDS (0.54, 95% CI: 0.44-0.64; p<0.0001), mREMS (0.62, 95% CI: 0.52-0.72; p<0.001) and SOFA (0.77, 95% CI: 0.69-0.85; p=0.003). On the other hand, the model retained a specificity that was comparable to that of internal medicine physicians (0.74, 95% CI: 0.64-0.82; p=0.509), abbMEDS (0.72, 95% CI: 0.64-0.81; p=0.327) and SOFA (0.74, 95% CI: 0.65-0.82, p=0.447), while still outperforming mREMS (0.64, 95% CI: 0.55-0.74; p=0.02). Additionally, the model had higher overall diagnostic accuracy with an AUC of 0.852 (95% CI: 0.783-0.922) compared to abbMEDS (0.631, 0.537-0.726, p=0.021), mREMS (0.630, 0.535-0.724, p=0.016), SOFA (0.752, 0.667-0.836, p=0.042) and internal medicine physicians (0.735, 0.648-0.821, p=0.032-0.189) (S4 Fig and S5 Table). Similar observations were made in additional evaluation metrics such as positive predictive value (NPV), negative predictive value (NPV) and accuracy (S5 Table). Individually, consultants were found to be more sensitive compared to residents (S5 Fig) with a poor to moderate agreement between the internists (Cohen's Kappa 0.46 to 0.67) (S6 Table). A sensitivity analysis with four additional comparisons, where one physician was excluded at a time, confirmed that the results are robust and that the outperformance of the machine learning model was not due to an outlier in the physician group (S7 Table).

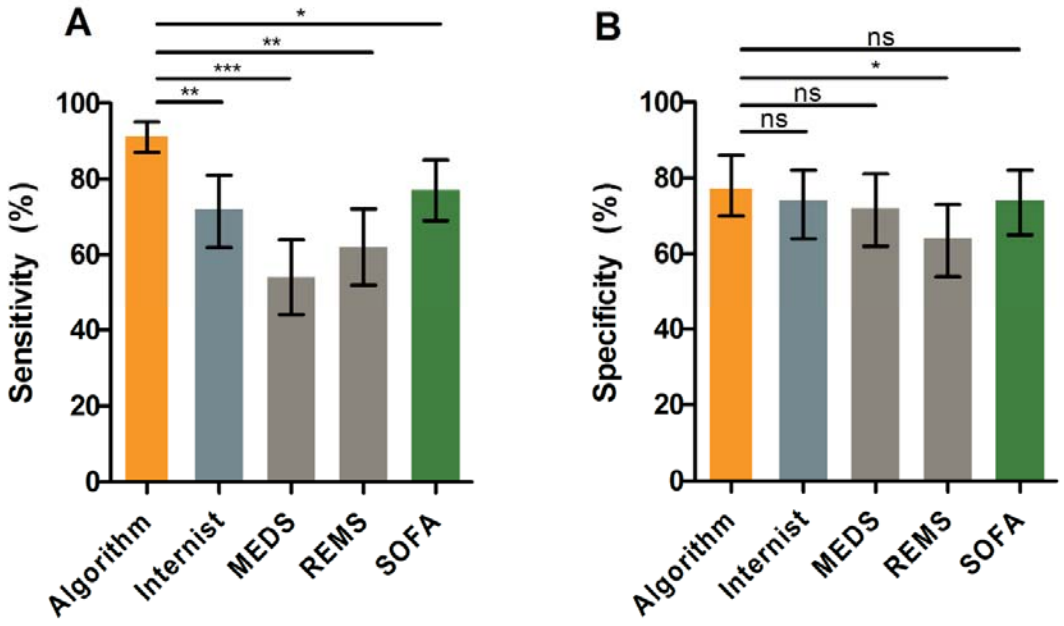

Fig 4. Comparison of XGBoost model with internal medicine physicians and clinical risk scores. The XGBoost model achieved a sensitivity (A) of 0.92 (95% CI: 0.87-0.95) and specificity (B) of 0.78 (95% CI: 0.70-0.86) for predicting mortality. This was significantly better than the mean prediction of internal medicine physicians for sensitivity (0.72, 0.62-0.81; p<0.001) as well as abbMEDS (0.54, 0.44-0.64; p<0.0001), mREMS (0.62, 0.52-0.72; p<0.001) and SOFA (0.77, 95% CI: 0.69-0.85; p=0.003). In terms of specificity, internal medicine physicians (0.74, 0.64-0.82; p=0.509), abbMEDS (0.72, 0.64-0.81; p=0.327) and SOFA (0.74, 95% CI: 0.65-0.82, p=0.447) achieved similar performance compared to the XGBoost model, opposed to mREMS (0.64, 0.55-0.74; p=0.02) which was significantly worse than machine learning predictions.

* = p<0.05; ** = p<0.001; *** = p<0.0001; NS = not significant.

Discussion

In the present study we demonstrate the application of machine learning models to predict 31-day mortality patients presenting to the ED with sepsis. Our study reports several important findings.

First, we show that machine learning based models can accurately predict 31-day mortality in patients with sepsis. Highest diagnostic accuracy was obtained with the model that was trained with both laboratory and clinical data. Patient characteristics that are employed in traditional risk scores, such as blood pressure and heart rate, were also found to be amongst the most important variables for model predictions. Second, machine learning models outperformed the judgment of internal medicine physicians and commonly used clinical risk scores, abbMEDS, mREMS and SOFA. Specifically, machine learning was more sensitive compared with risk scores and internal medicine physicians, while retaining identical or slightly higher specificity. These preliminary data provide support in favor of the development and implementation of machine learning based models as clinical decision support tools, e.g. risk stratification of sepsis patients presenting to the ED.

We are aware of several studies which describe the machine-learning based prediction of mortality in sepsis populations presenting to the ED [15-17]. Taylor et al. described a random forest model outperforming clinical risk scores in an ED population. Despite their bigger population, our XGBoost model appears to achieve similar performance to their random forest model, which corroborates and extends the power of this machine learning technique. Two recent studies by Barnaby et al. and Chiew et al. focused on using heart rate variability (HRV) for risk prediction in sepsis patients and reported predictive performance similar to our findings [15, 16]. Interestingly, their populations were smaller and this would therefore also advocate the use of HRV in our models. Despite these findings, Chiew et al. demonstrated that models without laboratory data significantly decreased in performance, emphasizing the importance of laboratory data in these machine learning models. Nevertheless, to the best of our knowledge this is the first study to report the direct comparison of machine learning models with internal medicine physicians. Although we do not present prospective results, we demonstrate that machine learning outperforms

clinical judgment of internal medicine physicians and clinical risk scores, implying that current XGBoost models potentially aid in risk stratification of ED patients. As an example, implementation of these models should revolve around identifying patients with a high risk, e.g. ≥50% mortality within 31 days, which would then be re-evaluated once more before being discharged from the ED. This kind of implementation was shown in a recent randomized clinical trial by Shimabukuro et al. [41], proving that average length of stay and in-hospital mortality decreased by using a ML-based sepsis detection model in the ICU. Although this was carried out with a small population in an ICU instead of the ED, it clearly shows the potential of ML-based risk stratifying models.

The current study has several strengths and limitations. Strengths include (i) comparison of laboratory versus laboratory + clinical models, (ii) analysis of features contributing to models' prediction and (iii) the comparison with internal medicine specialists. We are also aware of several limitations. First, the present study was a single-center study with a relatively small sample size at least from a machine learning analysis perspective. Nearly all machine learning models scale exceptionally well with data, and therefore substantial further improvement of diagnostic accuracy is likely when increasing the sample size. We also limited ourselves to sepsis patients presenting to the ED, and thus it is unknown to what degree these models translate to a broader, general ED population. Second, results presented in this study are based on retrospective data in a single center, limiting the external validity of the model. Unfortunately, this limitation currently applies to most studies applying ML in medicine. Third, the present study focused on model development and subsequent performance comparison with clinical judgment and clinical risk scores. It should be noted that the comparison with internal medicine specialists was performed using retrospectively generated electronic health records, rather than a prospective evaluation, which might have underestimated their diagnostic performance as they were not able to directly "see" the patient. Prospective evaluation, in respect to mortality, but also in relation to clinical endpoints that confirm true clinical benefit would facilitate implementation of ML-based risk stratification tools in clinical practice.

Conclusion

In conclusion, the present proof-of-concept study demonstrates the potential of machine learning models to predict mortality in patients with sepsis presenting to the ED. Machine learning outperformed clinical judgment of internal medicine physicians and established clinical risk scores. These data provide support in favor of the implementation of machine learning based risk stratification tools of sepsis patients presenting to the ED.

Acknowledgements

Not applicable.

References

1. LaCalle E, Rabin E. Frequent users of emergency departments: the myths, the data, and the policy implications. Ann Emerg Med. 2010;56(1):42-8. Epub 2010/03/30. doi: 10.1016/j.annemergmed.2010.01.032. PubMed PMID: 20346540.

2. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801-10. Epub 2016/02/24. doi: 10.1001/jama.2016.0287. PubMed PMID: 26903338; PubMed Central PMCID: PMCPMC4968574.

3. Roest AA, Tegtmeier J, Heyligen JJ, Duijst J, Peeters A, Borggreve HF, et al. Risk stratification by abbMEDS and CURB-65 in relation to treatment and clinical disposition of the septic patient at the emergency department: a cohort study. BMC Emerg Med. 2015;15:29. Epub 2015/10/16. doi: 10.1186/s12873-015-0056-z. PubMed PMID: 26464225; PubMed Central PMCID: PMCPMC4605126.

4. McLymont N, Glover GW. Scoring systems for the characterization of sepsis and associated outcomes. Ann Transl Med. 2016;4(24):527. Epub 2017/02/06. doi: 10.21037/atm.2016.12.53. PubMed PMID: 28149888; PubMed Central PMCID: PMCPMC5233540.

5. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):762-74. Epub 2016/02/24. doi: 10.1001/jama.2016.0288. PubMed PMID: 26903335; PubMed Central PMCID: PMCPMC5433435.

6. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Medicine. 1996;22(7):707-10. doi: 10.1007/BF01709751.

7. Olsson T, Terent A, Lind L. Rapid Emergency Medicine score: a new prognostic tool for in-hospital mortality in nonsurgical emergency department patients. J Intern Med. 2004;255(5):579-87. Epub 2004/04/14. doi: 10.1111/j.1365-2796.2004.01321.x. PubMed PMID: 15078500.

8. Vorwerk C, Loryman B, Coats TJ, Stephenson JA, Gray LD, Reddy G, et al. Prediction of mortality in adult emergency department patients with sepsis. Emerg Med J. 2009;26(4):254-8. Epub 2009/03/25. doi: 10.1136/emj.2007.053298. PubMed PMID: 19307384.

9. Crowe CA, Kulstad EB, Mistry CD, Kulstad CE. Comparison of severity of illness scoring systems in the prediction of hospital mortality in severe sepsis and septic shock. J

Emerg Trauma Shock. 2010;3(4):342-7. Epub 2010/11/11. doi: 10.4103/0974-2700.70761. PubMed PMID: 21063556; PubMed Central PMCID: PMCPMC2966566.

10. Olsson T, Terent A, Lind L. Rapid Emergency Medicine Score can predict long-term mortality in nonsurgical emergency department patients. Acad Emerg Med. 2004;11(10):1008-13. Epub 2004/10/07. doi: 10.1197/j.aem.2004.05.027. PubMed PMID: 15466141.

11. Minne L, Abu-Hanna A, de Jonge E. Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. Crit Care. 2008;12(6):R161. Epub 2008/12/19. doi: 10.1186/cc7160. PubMed PMID: 19091120; PubMed Central PMCID: PMCPMC2646326.

12. Rohacek M, Nickel CH, Dietrich M, Bingisser R. Clinical intuition ratings are associated with morbidity and hospitalisation. Int J Clin Pract. 2015;69(6):710-7. Epub 2015/02/18. doi: 10.1111/ijcp.12606. PubMed PMID: 25689155; PubMed Central PMCID: PMCPMC5024066.

13. Zelis N, Mauritz AN, Kuijpers LIJ, Buijs J, de Leeuw PW, Stassen PM. Short-term mortality in older medical emergency patients can be predicted using clinical intuition: A prospective study. PLoS One. 2019;14(1):e0208741. Epub 2019/01/03. doi: 10.1371/journal.pone.0208741. PubMed PMID: 30601815; PubMed Central PMCID: PMCPMC6314634.

14. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. Epub 2019/01/09. doi: 10.1038/s41591-018-0300-7. PubMed PMID: 30617339.

15. Barnaby DP, Fernando SM, Herry CL, Scales NB, Gallagher EJ, Seely AJE. Heart Rate Variability, Clinical and Laboratory Measures to Predict Future Deterioration in Patients Presenting With Sepsis. Shock. 2019;51(4):416-22. Epub 2018/05/31. doi: 10.1097/SHK.0000000000001192. PubMed PMID: 29847498.

16. Chiew CJ, Liu N, Tagami T, Wong TH, Koh ZX, Ong MEH. Heart rate variability based machine learning models for risk prediction of suspected sepsis patients in the emergency department. Medicine (Baltimore). 2019;98(6):e14197. Epub 2019/02/09. doi: 10.1097/MD.0000000000014197. PubMed PMID: 30732136; PubMed Central PMCID: PMCPMC6380871.

17. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. Acad Emerg Med. 2016;23(3):269-78. Epub 2015/12/19. doi: 10.1111/acem.12876. PubMed PMID: 26679719; PubMed Central PMCID: PMCPMC5884101.

18. Perng JW, Kao IH, Kung CT, Hung SC, Lai YH, Su CM. Mortality Prediction of Septic Patients in the Emergency Department Based on Machine Learning. J Clin Med. 2019;8(11).

Epub 2019/11/11. doi: 10.3390/jcm8111906. PubMed PMID: 31703390; PubMed Central PMCID: PMCPMC6912277.

19. Fagerstrom J, Bang M, Wilhelms D, Chew MS. LiSep LSTM: A Machine Learning Algorithm for Early Detection of Septic Shock. Sci Rep. 2019;9(1):15132. Epub 2019/10/24. doi: 10.1038/s41598-019-51219-4. PubMed PMID: 31641162; PubMed Central PMCID: PMCPMC6805937.

20. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open. 2018;8(1):e017833. Epub 2018/01/29. doi: 10.1136/bmjopen-2017-017833. PubMed PMID: 29374661; PubMed Central PMCID: PMCPMC5829820.

21. Klug M, Barash Y, Bechler S, Resheff YS, Tron T, Ironi A, et al. 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(1):220-7. Epub 2019/11/05. doi: 10.1007/s11606-019-05512-7. PubMed PMID: 31677104.

22. Sahni N, Simon G, Arora R. Development and Validation of Machine Learning Models for Prediction of 1-Year Mortality Utilizing Electronic Medical Record Data Available at the End of Hospitalization in Multicondition Patients: a Proof-of-Concept Study. J Gen Intern Med. 2018;33(6):921-8. Epub 2018/02/01. doi: 10.1007/s11606-018-4316-y. PubMed PMID: 29383551; PubMed Central PMCID: PMCPMC5975145.

23. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLoS One. 2017;12(4):e0174708. Epub 2017/04/07. doi: 10.1371/journal.pone.0174708. PubMed PMID: 28384212; PubMed Central PMCID: PMCPMC5383046.

24. Ford DW, Goodwin AJ, Simpson AN, Johnson E, Nadig N, Simpson KN. A Severe Sepsis Mortality Prediction Model and Score for Use With Administrative Data. Crit Care Med. 2016;44(2):319-27. Epub 2015/10/27. doi: 10.1097/CCM.0000000000001392. PubMed PMID: 26496452; PubMed Central PMCID: PMCPMC4724863.

25. Shukeri W, Ralib AM, Abdulah NZ, Mat-Nor MB. Sepsis mortality score for the prediction of mortality in septic patients. J Crit Care. 2018;43:163-8. Epub 2017/09/14. doi: 10.1016/j.jcrc.2017.09.009. PubMed PMID: 28903084.

26. Bogle B, Balduino, Wolk D, Farag H, Kethireddy, Chatterjee, et al. Predicting Mortality of Sepsis Patients in a Multi-Site Healthcare System using Supervised Machine Learning2019.

27. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. arXiv e-prints [Internet]. 2016 March 01, 2016. Available from: https://ui.adsabs.harvard.edu/abs/2016arXiv160302754C.

28. Nanayakkara S, Fogarty S, Tremeer M, Ross K, Richards B, Bergmeir C, et al. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. PLoS Med. 2018;15(11):e1002709. Epub 2018/12/01. doi: 10.1371/journal.pmed.1002709. PubMed PMID: 30500816; PubMed Central PMCID: PMCPMC6267953 following competing interests: KR is director of IntelliHQ Pty Ltd, non-profit AI innovation centre for healthcare, connected with Gold Coast University Hospital. KR is owner and Chairman of K. J. Ross & Associates Pty. Ltd. (KJR), professional services firm specialising in IT risk management and assurance. 20% of KJR's work is in healthcare. There is no direct financial stake in the results of the current study.

29. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, et al. 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. Crit Care Med. 2003;31(4):1250-6. Epub 2003/04/12. doi: 10.1097/01.CCM.0000050454.01978.3B. PubMed PMID: 12682500.

30. World Medical A. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013;310(20):2191-4. Epub 2013/10/22. doi: 10.1001/jama.2013.281053. PubMed PMID: 24141714.

31. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2(10):749-60. Epub 2019/04/20. doi: 10.1038/s41551-018-0304-0. PubMed PMID: 31001455; PubMed Central PMCID: PMCPMC6467492.

32. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. Explainable AI for Trees: From Local Explanations to Global Understanding. arXiv e-prints [Internet]. 2019 May 01, 2019. Available from: https://ui.adsabs.harvard.edu/abs/2019arXiv190504610L.

33. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020. doi: 10.1038/s42256-019-0138-9.

34. Lipovetsky S, Conklin M. Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry. 2001;17(4):319-30. doi: 10.1002/asmb.446.

35. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems. 2013;41:647-65.

36. Chang SH, Hsieh CH, Weng YM, Hsieh MS, Goh ZNL, Chen HY, et al. Performance Assessment of the Mortality in Emergency Department Sepsis Score, Modified Early Warning Score, Rapid Emergency Medicine Score, and Rapid Acute Physiology Score in Predicting Survival Outcomes of Adult Renal Abscess Patients in the Emergency

Department. Biomed Res Int. 2018;2018:6983568. Epub 2018/10/18. doi: 10.1155/2018/6983568. PubMed PMID: 30327779; PubMed Central PMCID: PMCPMC6169207.

37. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proceedings of the 22nd international conference on Machine learning; Bonn, Germany. 1102430: ACM; 2005. p. 625-32.

38. BRIER GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Monthly Weather Review. 1950;78(1):1-3. doi: 10.1175/1520-0493(1950)078<0001:Vofeit>2.0.Co;2.

39. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-45. Epub 1988/09/01. PubMed PMID: 3203132.

40. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82. Epub 2012/10/25. PubMed PMID: 23092060; PubMed Central PMCID: PMCPMC3900052.

41. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res. 2017;4(1):e000234. Epub 2018/02/13. doi: 10.1136/bmjresp-2017-000234. PubMed PMID: 29435343; PubMed Central PMCID: PMCPMC5687546.

Supporting information

S1 supporting information. Extended description of clinical criteria and risk scores.
S2 supporting information. Background information on machine learning models reviewed in the current study.

S1 Table. Overview of variables present in the datasets. The laboratory dataset consisted exclusively of laboratory variables with age, sex and time of request. The laboratory and clinical dataset contained all variables from the laboratory dataset and additionally clinical and vital characteristics.

S2 Table. Comparison of baseline statistical and machine learning models for predicting 31-day mortality risk. We performed a baseline comparison of statistical and machine learning models (S1 supporting information) for the 31-day mortality prediction task using the laboratory dataset. We used five-fold cross validation to assess model performance. Performance was assessed by area under the receiver operating characteristic curve (AUC) and accuracy. Confidence intervals were calculated using bootstrapping methods (n=1,000).

S3 Table. Hyperparameters of XGBoost models. Hyperparameters were based on theoretical reasoning rather than hyperparameter tuning. This was done to prevent overfitting on hyperparameters due to small sample size. “Base_score”, “Missing”, “Reg_alpha”, “Reg_lambda” and “Subsample” parameters were standard values provided by the XGBoost interface. “Max_depth”, “max_delta_step” and “estimators” were values we internally use for these kind of machine learning models. During the study, hyperparameters were never adjusted to gain performance in our validation dataset.

S4 Table. Extended analysis of correlation between important model features and clinical risk scores. To study the correlation between the most important features contributing to model predictions and the clinical criteria (qSOFA and SIRS) and risk scores (abbMEDS and mREMS), we compared their existence in both. The top-20 most important features (Fig 3 in main article) are compared to all criteria in the clinical scores (S1 supporting information). We observe that most of the features present in the clinical criteria and scores are also among the most important features in the lab and clinical machine learning model.

S5 Table. Extended comparison of machine learning models with internal medicine physicians and clinical risk scores. In addition to sensitivity and specificity, we evaluated the performance of each group by positive predictive value (PPV), negative predictive value (NPV), accuracy and area-under-the receiver operating characteristics curve (AUC). Our

XGBoost model shows superior performance in each of these metrics, which is in line with the findings presented in the manuscript.

S6 Table. Inter-rater agreement of internal medicine physicians. Cohen's kappa was used to measure the inter-rater agreement between the internal medicine physicians. The level of agreement was interpreted as nil if κ was 0 to 0.20; minimal, 0.21 to 0.39; weak, 0.40 to 0.59; moderate, 0.60 to 0.79; strong, 0.80 to 0.90; and almost perfect, 0.90 to 1.3.

S7 Table. Machine learning comparison to alternating physician groups. In each comparison between the machine learning model and the physicians group, a single physician was removed from the physician group. In every comparison the machine learning model outperforms the physicians. This analysis shows that the higher performance of the machine learning model was not due to systemic underperformance of a single physician.

S1 Fig. Flow diagram of study inclusion. During the study period 5,967 patients that presented to our emergency department were referred to an internal medicine physician. Of these patients, 1420 patients fulfilled two or more SIRS and/or qSOFA criteria. After exclusion of 76 patients, a number of 1,344 patients were separated into development and validation datasets.

S2 Fig. Five-fold cross validation of diagnostic performance of XGBoost models. During each cycle of cross-validation, we assessed predictive performance by area under the receiver operating characteristic curves (AUC). Performance was determined for models trained with laboratory data (A) and models trained with laboratory and clinical data (B) to predict 31-day mortality.

S3 Fig. Five-fold cross validation of calibration of XGBoost models. During each cycle of cross-validation, we assessed calibration by calibration curves and their respective brier scores. Calibration was determined for models trained with laboratory data (A) and models trained with laboratory and clinical data (B).

S4 Fig. Receiver operating characteristic analysis of machine learning model, risk scores and internal medicine physicians. Receiver operating characteristics analysis of the lab + clinical machine learning model (AUC: 0.852 [0.783-0.922]), abbMEDS (0.631 [0.537-0.726]), mREMS (0.630 [0.535-0.724]) and internal medicine physicians (mean 0.735 [0.648-0.821]). Internal medicine physicians were depicted as bullets in the ROC analysis.

S5 Fig. Individual performance of internal medicine physicians. Predictive performance of all internal medicine specialists (n=4; 2 experienced consultants in acute internal medicine and 2 experienced residents acute internal medicine) was assessed by sensitivity (left) and specificity (right). Consultants (experienced) specialists are depicted in grey and residents in orange.