Transparent Machine Learning models for Rapid Risk Stratification in the Emergency Department: A multi-center evaluation

William P.T.M. van Doorn[1,2], Floris Helmich[3], Paul M.E.L. van Dam[4], Leo H.J. Jacobs[5], Patricia M. Stassen[4,6], Otto Bekers[1,2], Steven J.R. Meex[*1,2]

Affiliations

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

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

[3] Department of Clinical Chemistry & Hematology, Zuyderland Medical Center, Heerlen, The Netherlands.

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

[5] Laboratory of Clinical Chemistry, Meander Medical Center, Amersfoort, The Netherlands.

[6] 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: 268

Word count: 2967

Table and Figures: 4

References: 50

Abstract

Introduction: Risk stratification of patients presenting to the emergency department (ED) is important for appropriate triage. Using machine learning technology, we can integrate laboratory data from a modern emergency department and present these in relation to clinically relevant endpoints for risk stratification. In this study, we developed and evaluated transparent machine learning models in four large hospitals in the Netherlands.

Methods: Historical laboratory data (2013-2018) available within the first two hours after presentation to the ED of Maastricht University Medical Centre+ (Maastricht), Meander Medical Center (Amersfoort), and Zuyderland (locations Sittard and Heerlen) were used. We used the first five years of data to develop the model and the sixth year to evaluate model performance in each hospital separately. Performance was assessed using area under the receiver-operating-characteristic curve (AUROC), brier scores and calibration curves. The SHapley Additive exPlanations (SHAP) algorithm was used to obtain transparent machine learning models.

Results: We included 266,327 patients with more than 7 million laboratory results available for analysis. Models possessed high diagnostic performance with AUROCs of 0.94 [0.94-0.95], 0.98 [0.97-0.98], 0.88 [0.87-0.89] and 0.90 [0.89-0.91] for Maastricht, Amersfoort, Sittard and Heerlen, respectively. Using the SHAP algorithm, we visualized patient characteristics and laboratory results that drive patient-specific $RISK^{INDEX}$ predictions. As an illustrative example, we applied our models in a triage system for risk stratification that categorized 94.7% of the patients as low risk with a corresponding NPV of ≥99%.

Discussion: Developed machine learning models are transparent with excellent diagnostic performance in predicting 31-day mortality in ED patients across four hospitals. Follow up studies will assess whether implementation of these algorithm can improve clinically relevant endpoints.

Introduction

An increasing number of patients are referred to emergency departments (ED) worldwide [1-3]. Prolonged waiting times and associated crowding have been shown to increase mortality [4] and rapid risk stratification is, therefore, a core task in emergency medicine. Identifying patients at high and low risk shortly after admission could help decision-making in the prioritization of patients, treatment, level of observation, and post-discharge follow-up. Numerous traditional risk scores and triage systems for stratification of patients in the ED are available, such as the modified early warning score (MEWS), rapid emergency medicine score (REMS) and emergency severity index (ESI) [5-9]. Unfortunately, these systems often generalize poorly and lack precision, which impedes their use on a patient level [10, 11].

Modern emergency departments generate vast amounts of clinical, physical and laboratory data. This data is generally heterogeneous and comprises both structured and unstructured information. Machine learning allows the integration of these data on a human interpretable level in relation to clinically relevant endpoints. Recently, machine-learning based mortality prediction models were developed using data extracted from patients in the ED [12-19]. Although these models were superior to traditional risk scores and physicians [12, 15, 17, 20], most of are perceived to be so-called “black boxes”, which not only could limit their acceptance among clinicians but also raise legal and ethical concerns. Models explaining patient-specific predictions have emerged, which might increase the understanding of, and trust in, machine learning prediction models [21-24]. This could, in turn, facilitate the translation and acceptance of machine learning models into clinical decision-support tools.

In this study, we used machine learning technology in four hospitals to develop local, transparent machine learning models to accurately predict 31-day mortality risk. We aimed to provide an individual assessment of a patient’s mortality risk (**the RISK^{INDEX} score**), which is an illustrative example of a clinical decision support system, using baseline patient characteristics and laboratory data.

Methods

Study design and setting

We performed a multi-center, retrospective cohort study among all patients who presented to the ED at the Maastricht University Medical Center (Maastricht, The Netherlands), Meander Medical Center (Amersfoort, The Netherlands) and Zuyderland medical Center locations Sittard (Sittard, The Netherlands) and Heerlen (Heerlen, The Netherlands) between January 1, 2013 and December 31, 2018. For convenience, we will refer to each of the centers by their respective location; Maastricht, Amersfoort, Sittard and Heerlen. This study was approved by the medical ethical committees of each of the individual centers (Maastricht: #2018-0838, Amersfoort: TWO19-46, Sittard: #2018-0838, Heerlen: #2018-0838). The study follows the STROBE guidelines [25] and was conducted according to the principles of the Declaration of Helsinki [26].

Patient population

We included all patients presenting to the ED aged ≥18 years with at least 3 laboratory tests ordered by the attending physician. Patients whose previous presentation to the ED was less than 48 hours ago were excluded.

Dataset construction

Data anonymization, collection, processing, model selection and development were performed for each of the four hospitals separately. We collected all available laboratory data of the patients ordered within two hours after the first laboratory request from the ED. All laboratory data acquired after two hours were not used for model development. Rare laboratory tests requested in less than 1:10,000 patients of that hospital were excluded. The primary outcome measure for the study was mortality within 31 days after initial ED presentation and was acquired through the electronic health record.

Each hospital comprised 6 year of data from consecutive patients. The first five years were used for model development, the sixth year was completely retained from model development and used to evaluate the performance of developed machine learning models. The first five years of the dataset (model development) were randomly split into training (70%), tuning (20%) and calibration datasets (10%) such

that data from a given presentation was present in one split only. The training split was used to train the proposed models. The tuning set was used to iteratively improve the models by selecting the best model architectures and hyperparameters, and the calibration split was used to perform post-hoc calibration on the model predictions. Finally, the sixth year of the dataset, the validation dataset, was used to evaluate the performance of machine learning models.

Model selection, training and calibration

We aimed to develop a clinical decision support tool that uses heterogeneously requested laboratory results from patients presenting to the emergency department to predict the likelihood of 31-day mortality by generating a calibrated value between and 100. We termed this output score the $RISK^{INDEX}$. Ideally, this value directly translates to the probability of the patient dying within 31 days. Various statistical and machine learning algorithms can be applied to develop such a clinical decision support tool, including regression techniques [27, 28], neural network architectures [29, 30], gradient boosting systems [31-34] and decision trees [35] (Supplemental section A).

The light gradient boosting system (LightGBM) architecture was selected amongst several alternatives on the basis of the tuning set performance (Supplemental information section A and Table 1). LightGBM is an implementation of distributed, efficient gradient-boosting systems with native support for missing values [33]. Next, we evaluated a broad spectrum of hyperparameter combinations for this architecture (Supplemental Table 2). Hyperparameter optimization is the process of selecting a set of optimal hyperparameters, which are features controlling the training process of a machine learning model such as the rate of learning and the maximum level of complexity. We performed bayesian hyperparameter optimization using tree-parzen estimators (TPE) [36]. In bayesian hyperparameter optimization we build a probability model of the objective function and use it to select the most promising hyperparameters to evaluate in the true objective function. For this study, optimization was run for 1,000 iterations with logarithmic loss as our objective function. The search space was defined in Supplemental Table 2. Hyperparameter optimization resulted in LightGBM architectures consisting of 220 – 740 boosted trees with a maximum depth of 11 – 37 and maximum leaves of 320 – 690 for each

base learner (see Supplemental Table 3). We used exponential learning-rate decay during training. The best validation results were achieved using an initial learning rate of 0.075 – 0.145 decaying every 2,000 training steps by a factor of 0.7-0.8. The loss function during training was logarithmic loss.

LightGBM models with optimal hyperparameters were recalibrated on the calibration set in order to further improve the quality of the risk predictions. Recalibration was performed as LightGBM models are prone to miscalibration, essentially meaning that the generated $RISK^{INDEX}$ does not correlate with the true mortality chance. Hence, recalibration ensures that consistent probabilistic interpretations of the $RISK^{INDEX}$ predictions can be made [37]. For calibration, we considered Platt scaling [38], isotonic regression [39] and Platt-Binner scaling [40]. Model calibration was assessed by the Brier score [41] and visual inspection of reliability plots [42]. Reliability plots are the usual approach for evaluating calibration of binary outcomes in which we compare decile-binned means of predictions versus means of the observed outcomes in the patients. We used Platt-Binner scaling as this was shown to result in the best calibrated models (see Supplemental Figure 1). The resulting calibrated predictions were defined as the $RISK^{INDEX}$. Data preprocessing, model development, selection, training and calibration was performed using Python programming language (version 3.7.1) using packages Numpy (version 1.17), Pandas (version 0.24), Keras (version 2.2.2), scikit-learn (version 0.22.0) and tensorflow (version 2.0.1, beta).

Model evaluation

We evaluated overall model performance in the validation set by 1) area under the receiver-operating-characteristic curve (AUROC) to quantify the ability of models to discriminate between survivors and non-survivors, and 2) visual inspection of calibration curve and Brier scores to estimate how accurately $RISK^{INDEX}$ scores estimate the likelihood of 31-day mortality. Next, we created an embedded reference table based on the validation dataset to report estimates of sensitivity, negative predictive value (NPV), specificity and positive predictive value (PPV) for each $RISK^{INDEX}$ score between 0-100. This table was subsequently used to compare diagnostic metrics from the model (sensitivity, NPV, specificity, and PPV) across the four hospitals at certain selected statistical thresholds.

Model transparency

To explain the $RISK^{INDEX}$ generated by our machine learning models, we applied the Shapley additive explanations (SHAP) algorithm. SHAP allows us to obtain explanations of the patient characteristics and laboratory results (further referred to as "variables") that drive patient-specific predictions to mitigate the issue of black-box predictions. SHAP is a model-agnostic representation of feature importance where the impact of each variable on a particular prediction is represented using Shapley values inspired by cooperative game theory and their extensions [43-45]. 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 Shapley values contribute towards a negative prediction (survival).

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. We used 1,000 bootstrapped to calculate 95% confidence intervals, unless otherwise mentioned. Model evaluation and statistical analysis was performed using Python (version 3.7.1) using packages Numpy (version 1.17), Pandas (version 0.24) and Matplotlib (version 3.1.2).

Results

Patient and laboratory characteristics

In the current study we included more than 50,000 presentations for each hospital resulting in a total of 266,327 unique presentations to the ED. The total population consisted of slightly more female (mean; 50.8%) patients with a mean age of 61.5 (+-22.4) years. Within the first two hours of presentations, on average 29 (± 10.6) laboratory parameters were requested. Among these parameters there is some heterogeneity between centers, but complete blood count, electrolytes and lactate dehydrogenase (LD) were amongst the most prevalent in all hospitals. 31-Day mortality rates were 6.4%, 4.0%, 5.9% and 5.0% for Maastricht, Amersfoort, Sittard and Heerlen, respectively. Baseline characteristics are described in **Table 1**.

Table 1 | Baseline and laboratory characteristics of the four study populations.

	MUMC, Maastricht N = 66,770	**Meander, Amersfoort N = 81,152**	**Zuyderland, Sittard N = 66,423**	**Zuyderland, Heerlen N = 51,982**
Demographics				
Age, years	59 (± 22.2)	59 (± 23.7)	65 (± 21.3)	64 (± 21.5)
Sex, female (%)	32,829 (49.2%)	41,907 (51.6%)	34,256 (51.6%)	26,185 (50.3%)
Laboratory				
Mean number of tests per patient	22 (± 10.0)	25 (± 8.0)	31 (± 10.4)	31 (± 10.8)
Ten most frequent laboratory orders, n (%)	Creatinine, 59,937 (89.9%) CBC[a], 59,632 (89.4%) Sodium, 56,529 (84.8%), Potassium, 56,487 (84.7%) CRP[b], 55,178 (82.7%) Urea, 53,972 (80.9%) Platelets, 47,059 (70.6%) Glucose, 43,733 (65.6%) ALAT[c], 36.429 (54.6%) ASAT[d], 32.130 (48.1%)	CBC[a], 77,625 (95.7%) CRP[b], 76,639 (94.4%) Sodium, 75,638 (93.2%) Creatinin, 75,393 (92.9%) Potassium, 75,025 (92.4%) Glucose, 75,018 (92.4%) Urea, 72.115 (88.9%) CK, 64.356 (79.3%) Platelets, 54.380 (67.0%) ALAT[c], 54.279 (66.9%)	CBC[a], 62,518 (94.1%) Platelets, 62,517 (94.1%) CRP[b], 60,910 (91.7%) Creatinin, 60,046 (90.4%) Sodium, 58,558 (88.1%) Potassium, 58,404 (87.9%) Glucose, 57,987 (87.3%) Urea, 57,559 (86.6%) ALAT[c], 53,968 (81.2%) ASAT[d], 53.914 (81.2%)	CBC[a], 49,056 (94.4%) Platelets, 49,040 (94.3%) CRP[c], 47,573 (91.5%) Creatinin, 47,043 (90.5%) Sodium, 46,615 (89.7%) Potassium, 46,364 (89.2%) Glucose, 45,536 (87.6%) Urea, 45,074 (86.7%) ALAT[c], 42,486 (81.7%) ASAT[d], 42.415 (81.6%)
Outcome				
31-day mortality	4,242 (6.4%)	3,277 (4.0%)	3,917 (5.9%)	2,603 (5.0%)

[a] **Complete Blood Count including hemoglobin, hematocrit, MCH, MCV and white blood cells;** [b] **C-reactive Protein;** [c] **Alanine (Amino)Transaminase;** [d] **Aspartate (Amino)Transaminase**

Model performance

We developed machine learning models that predict the 31-day mortality likelihood of an individual patient presenting to the ED: the RISK^{INDEX}. Machine learning models were able to discriminate between patients who died or survived within 31-days as depicted by AUROCs of 0.944 [0.935-0.951], 0.978 [0.973-0.982], 0.877 [0.866-0.888] and 0.904 [0.894-0.914] for Maastricht, Amersfoort, Sittard and Heerlen, respectively **(Figure 1A)**. After calibration (see supplemental Figure 1), the RISK^{INDEX} correlated well with actual mortality frequency (**Figure 1B)**. Hence, the RISK^{INDEX} provides an individualized and precise assessment of 31-day mortality risk by combining available laboratory data and patient characteristics requested within the first two hours after presentation.

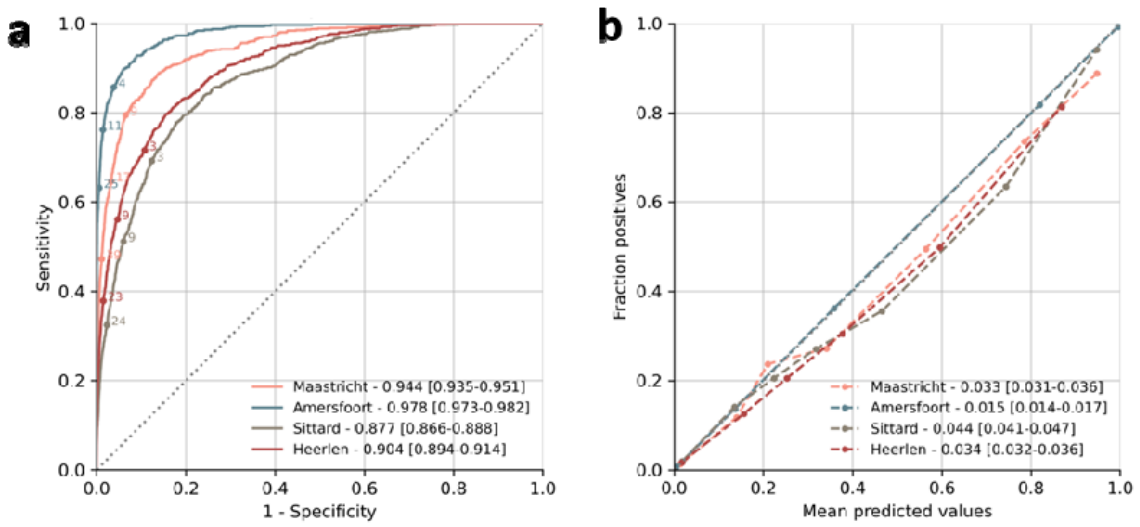

Figure 1 | Diagnostic performance and calibration of machine learning models. (A) Receiver operating characteristic curves (ROC) showing the discrimination of the LightGBM models in each of the different centers. Annotated points depict example RISK^{INDEX} thresholds for illustrative purposes. (B) Calibration of the machine learning models with the observed proportion of 31-day mortality in each of the centers. Each point represents 10% of the patients in the validation dataset.

Proposal for clinical decision support tool

The RISK^{INDEX} is by design a continuous measure, with a high RISK^{INDEX} translating to a high likelihood of 31-day mortality and low RISK^{INDEX} translating to low 31-day mortality risk (see Supplementary Figures 1-4). In clinical practice however, most clinical decision support tools use fixed thresholds to categorize patients as low, medium or high risk. Our RISK^{INDEX} can be easily transformed to such a fixed-threshold decision support tool, and users can adjust thresholds to their desired level of risk tolerance. An illustrative example of how an individual hospital may employ the RISK^{INDEX} is as follows: define the acceptable percentage of patients that are erroneously identified as “low-risk” by the algorithm in your emergency department (any number from 0-100%). This percentage, e.g. 1%, could be derived from an inventory of acceptable risk tolerance for adverse events by patients, health care workers, or both [46]. Then, use the corresponding negative predictive value (in this case 99%) to derive the matching RISK^{INDEX} threshold from the calibration set and associated values for sensitivity, specificity, and proportion of subjects identified as low risk (**Table 2**). A similar approach can be applied to identify high risk patients: define the positive predictive value that would provide an acceptable balance between true high risk patient identification and false positives, e.g. a positive predictive value of 75% would categorize between 1.1% and 3.9% as high-risk individuals with 1 in 4 “flaggings” by the clinical decision support tool being false positive **(Table 2).** A higher proportion of high risk subject identification is feasible but will be at the expense of increased false positive flaggings.

Table 2: Illustrative example of clinical decision support tool using developed machine learning models. The fixed-threshold decision support tool was fixed at a negative predictive value of 99% to identify low-risk patients (corresponding $RISK^{INDEX}$ cut-offs between 1.0 and 12.1) and fixed at a positive predictive value of 75% to identify high-risk patients (corresponding $RISK^{INDEX}$ cut-offs between 29.9 and 64). Diagnostic metrics and proportion of patients identified as either low- or high-risk are described in the table.

Low-risk defined at a negative predictive value of 99%					
	$RISK^{INDEX}$ cut-off	**Sensitivity**	**Specificity**	**NPV**	**Proportion of patients**
Maastricht	4.3	86.9% [84.2% - 88.9%]	87.7% [87.2% - 88.2%]	99.0% [98.8% - 99.2%]	83.0% [82.3% - 83.6%]
Meander	12.1	75.9% [72.3% - 78.6%]	97.5% [97.3% - 97.7%]	99.0% [98.8% - 99.1%]	94.7% [94.3% - 95.0%]
Sittard	1.0	85.5% [83.2% - 87.3%]	75.0% [74.2% - 75.7%]	98.8% [98.6% - 99.0%]	71.5% [70.7% - 72.0%]
Heerlen	1.0	82.7% [80.0% - 85.2%]	81.2% [80.8% - 81.8%]	98.9% [98.8% - 99.1%]	78.1% [77.7% - 78.8%]
High-risk defined at positive predictive value of 75%					
	$RISK^{INDEX}$ cut-off	**Specificity**	**Specificity**	**PPV**	**Proportion of patients**
Maastricht	41.1	45.7% [42.3% - 49.5%]	99.0% [98.8% - 99.1%]	74.9% [70.8% - 77.8%]	3.9% [3.5% - 4.2%]
Meander	29.9	60.6% [56.9% - 64.1%]	99.2% [99.0 - 99.3]	74.9% [70.8% - 78.3%]	3.1% [2.9% - 3.4%]
Sittard	64	14.7% [12.6% - 16.7%]	99.7% [99.6% - 99.8%]	74.8% [68.8% - 80.6%]	1.1% [1.0% - 1.3%]
Heerlen	41.1	27.1% [24.2% - 29.4%]	99.5% [99.4% - 99.7%]	74.8% [70.6% - 80.6%]	1.7% [1.5% - 2.0%]

Transparent model predictions

In order to obtain transparent machine learning models, it is necessary to be able to explain the $RISK^{INDEX}$ generated by our machine learning models. Hence, we applied the SHAP algorithm to map the importance of patient characteristics and laboratory results to the generated $RISK^{INDEX}$ (see Supplemental Figures 6-8). This is illustrated for a low-, medium and high-risk individual in **Figure 2**. For example, a high $RISK^{INDEX}$ was generated for a 71-year old (+ 5.5 $RISK^{INDEX}$) individual with a high numeric amount of laboratory measurements (+ 31.5 $RISK^{INDEX}$), a high lactate dehydrogenase (LD; +17.7) and a low albumin level (+6), whilst the remainder of the features contributed another significant portion (+10.6) ultimately leading to a $RISK^{INDEX}$ of 76. On the other hand, the low-risk individual had a normal albumin level (- 0.9 $RISK^{INDEX}$), the presence of a pH blood measurement (-0.6), a normal lymphocyte level (-0.5) and the remainder of the features which also lowered the prediction (-1.3). Yet, a relatively high urea level (17 mmol/L) still caused a small increase in $RISK^{INDEX}$ (+0.7).

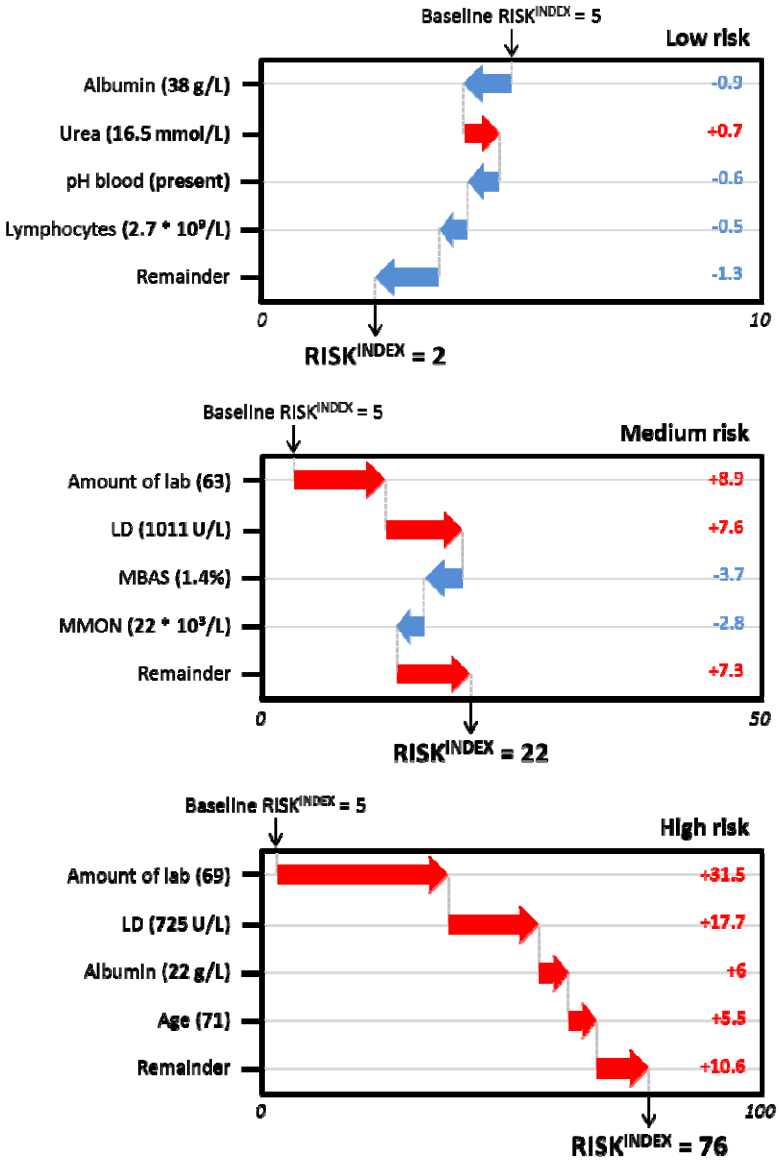

Figure 2: Model explanation in low-, medium- and high-risk individuals. In order to obtain transparent machine learning models, it is necessary to explain the RISKINDEX generated by our machine learning models. Here we illustrate the importance of patient characteristics and laboratory tests for a low (upper), medium (middle) and high-risk (lower) individual with $RISK^{INDEX}$ scores of 5, 22 and 76, respectively.

Discussion

In a large, multi-center study of more than 260,000 patients presenting to the emergency department across four hospitals, we used machine learning to develop and evaluate a novel clinical decision support tool that incorporates baseline laboratory data available within two hours after presentation to accurately predict the probability of the patient dying within 31 days. We developed transparent machine learning models which were well calibrated and had overall high diagnostic performance. Our study has several unique characteristics.

First, our RISK^{INDEX} clinical decision support tool provides an individualized, precise and rapid assessment of 31-day mortality risk by using baseline laboratory results acquired within two hours after the presentation of the patient. Our models had high diagnostic performance with AUROCS of 0.944 [0.935-0.951], 0.978 [0.973-0.982], 0.877 [0.866-0.888] and 0.904 [0.894-0.914] (Maastricht, Amersfoort, Sittard and Heerlen), outperforming any clinical decision support tool or risk score currently used in the emergency department for risk stratification [6, 7, 47].

Second, we used the Shapley additive explanations (SHAP) algorithm to obtain transparent machine learning models [21, 23]. The SHAP algorithm allows us to visualize the importance of patient characteristics and laboratory results that drive patient-specific RISK^{INDEX} predictions **(as illustrated in Figure 2)**. Development of such transparent machine learning models mitigate the issue of “black-box” predictions, and contribute to the understanding and acceptance of these models amongst clinicians and nurses. Furthermore, transparency in these models will likely become inevitable as regulations already expressed their concern with black-box predictions, signaling that automated prediction systems are enforced to inform users about the logic involved, as well as the significance and the envisaged consequences of its predictions in the near future [48-50].

Third, our clinical decision support tool is very versatile as it can be adjusted to the demands of the specific healthcare system or institution. For example, we illustrate a triage algorithm using a negative predictive value of 99% to identify low-risk patients, and a positive predictive value of 75% to identify high-risk patients (**Table 2)**. Nevertheless, in a more conservative institution we can adjust the low-risk thresholds

accordingly, e.g. to an even higher NPV of 99.5% implying that only 5 out of a 1.000 patients would erroneously be identified as “low-risk”. Implementation of such a triage system using our proposed clinical decision tool is convenient as current models rely on data that are easily acquired through existing laboratory system infrastructure. This is an advantage compared to machine learning models trained with e.g. clinical data that require manual annotation or collection which is complicated to automate in a prospective, real-world setting.

Fourth, we show that the methodology is robust and consistent by developing a clinical decision support tool for each hospital separately. Differences in diagnostic performance between hospitals might in part be explained by the geographic nature of the hospital (rural versus urban), the baseline mortality rates, and the laboratory testing patterns of the attending physicians. It would be of particular interest to unravel the source of these differences in order to improve the model performance at a hospital level.

Fifth, the large sample size of more than 260,000 patients and 7.1 million laboratory tests allowed for the development of machine learning models with high performance. Despite our models being trained almost exclusively with laboratory data, they outperform machine learning models which also had full access to clinical data of a patient [18, 19]. This highlights that -regardless of having access to less data concerning an individual patient (e.g. no clinical characteristics)- sample size is extremely important in building high-performance machine learning models.

Literature

We are aware of numerous attempts to use machine learning technology for risk stratification in the ED in a retrospective setting [12-14, 18, 19]. Klug et al. and Perng et al. developed machine learning models with similar performance as presented in this study (AUCs of 0.96 and 0.93, respectively) [18, 19]. Although diagnostic performance was similar, there are some notable differences. First, these studies focused on populations from a single center which we extended by developing and evaluating models in four hospitals. Second, these studies used clinical and vital characteristics of patients whereas we almost exclusively relied on the laboratory results. Third, we provide explanations of our generated $RISK^{INDEX}$ scores on a

patient level using the recent SHAP algorithm. Fourth, we provide illustrative implementation strategies using pre-defined safety (NPV) and efficacy (PPV) measures to identify low- and high-risk patients at the emergency department, respectively.

Limitations

Several limitations should be recognized. First, the current study is based on retrospective data and prospective studies are warranted to study performance and true clinical benefit of our clinical decision support tool in a real-world setting. This would also allow us to study the (dis)advantages of implementing these models using a triage system based on statistical thresholds (**e.g. Figure 3**) compared to an approach based on individual $RISK^{INDEX}$ estimates. Second, mortality information was retrieved through the laboratory information system, which is not fully connected to the national person registry. As a result we most likely report an underestimation of true mortality rates. Third, these models possess -despite being explainable- algorithmic bias; models have been trained entirely upon the basis of what humans have done before. This implies that model predictions cannot be extrapolated, and that predictions in e.g. minority populations have a higher degree of uncertainty. To facilitate the interpretation of such uncertain predictions, it would be desirable to implement uncertainty measures amongst the prediction, e.g. in form of confidence intervals. These uncertainty measures could then warn clinicians when a certain prediction is highly uncertain, ultimately leading to increased trust amongst the users of these clinical decision support tools.

Conclusion

Our novel $RISK^{INDEX}$ clinical decision support tool incorporates patient characteristics and laboratory tests available within the first two hours after presentation to provide an individual, precise and transparent assessment of the patient's mortality risk within 31 days. These models had overall high diagnostic performance, are explainable, and can be implemented in a triage system extending current systems used in modern emergency departments. Prospective, follow-up studies are warranted to study the feasibility and performance of these models in a real-world clinical setting.

Acknowledgements:

None.

Funding:

This study was funded by a Noyons Stipendium from the Dutch Federation of Clinical Chemistry (NVKC).

Disclosures:

Nothing to declare in relation to the current manuscript.

References:

1. LaCalle, E. and E. Rabin, *Frequent users of emergency departments: the myths, the data, and the policy implications.* Ann Emerg Med, 2010. **56**(1): p. 42-8.
2. Hooker, E.A., P.J. Mallow, and M.M. Oglesby, *Characteristics and Trends of Emergency Department Visits in the United States (2010-2014).* J Emerg Med, 2019. **56**(3): p. 344-351.
3. Wansink, L., et al., *Trend analysis of emergency department malpractice claims in the Netherlands: a retrospective cohort analysis.* Eur J Emerg Med, 2019. **26**(5): p. 350-355.
4. Guttmann, A., et al., *Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada.* BMJ, 2011. **342**: p. d2983.
5. Seymour, C.W., 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): p. 762-74.
6. Olsson, T., A. Terent, and L. Lind, *Rapid Emergency Medicine Score can predict long-term mortality in nonsurgical emergency department patients.* Acad Emerg Med, 2004. **11**(10): p. 1008-13.
7. Vorwerk, C., et al., *Prediction of mortality in adult emergency department patients with sepsis.* Emerg Med J, 2009. **26**(4): p. 254-8.
8. Christ, M., et al., *Modern triage in the emergency department.* Dtsch Arztebl Int, 2010. **107**(50): p. 892-8.
9. Crowe, C.A., et al., *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): p. 342-7.
10. Collins, G.S., et al., *External validation of multivariable prediction models: a systematic review of methodological conduct and reporting.* BMC Med Res Methodol, 2014. **14**: p. 40.
11. Ha, D.T., et al., *Prognostic performance of the Rapid Emergency Medicine Score (REMS) and Worthing Physiological Scoring system (WPS) in emergency department.* Int J Emerg Med, 2015. **8**: p. 18.

12. Taylor, R.A., 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): p. 269-78.
13. Shafaf, N. and H. Malek, *Applications of Machine Learning Approaches in Emergency Medicine; a Review Article.* Arch Acad Emerg Med, 2019. **7**(1): p. 34.
14. Tang, F., et al., *Predictive modeling in urgent care: a comparative study of machine learning approaches.* JAMIA Open, 2018. **1**(1): p. 87-98.
15. Levin, S., et al., *Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index.* Ann Emerg Med, 2018. **71**(5): p. 565-574 e2.
16. Peck, J.S., et al., *Generalizability of a simple approach for predicting hospital admission from an emergency department.* Acad Emerg Med, 2013. **20**(11): p. 1156-63.
17. Barnes, S., et al., *Real-time prediction of inpatient length of stay for discharge prioritization.* J Am Med Inform Assoc, 2016. **23**(e1): p. e2-e10.
18. Klug, M., 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): p. 220-227.
19. Perng, J.W., et al., *Mortality Prediction of Septic Patients in the Emergency Department Based on Machine Learning.* J Clin Med, 2019. **8**(11).
20. Meex, v.D., *PLACEHOLDER SEPSIS.* 2020.
21. Lundberg, S.M., et al., *From local explanations to global understanding with explainable AI for trees.* Nature Machine Intelligence, 2020.
22. Lundberg, S.M., et al., *Explainable machine-learning predictions for the prevention of hypoxaemia during surgery.* Nat Biomed Eng, 2018. **2**(10): p. 749-760.
23. Thorsen-Meyer, H.-C., et al., *Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records.* The Lancet Digital Health, 2020.
24. Hyland, S.L., et al., *Early prediction of circulatory failure in the intensive care unit using machine learning.* Nat Med, 2020. **26**(3): p. 364-373.

25. von Elm, E., et al., *The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies.* Lancet, 2007. **370**(9596): p. 1453-7.
26. World Medical, A., *World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects.* JAMA, 2013. **310**(20): p. 2191-4.
27. Bagley, S.C., H. White, and B.A. Golomb, *Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain.* J Clin Epidemiol, 2001. **54**(10): p. 979-85.
28. Harrell, F.E., Jr., et al., *Regression models for prognostic prediction: advantages, problems, and suggested solutions.* Cancer Treat Rep, 1985. **69**(10): p. 1071-77.
29. Forsstrom, J.J. and K.J. Dalton, *Artificial neural networks for decision support in clinical medicine.* Ann Med, 1995. **27**(5): p. 509-17.
30. Agatonovic-Kustrin, S. and R. Beresford, *Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research.* J Pharm Biomed Anal, 2000. **22**(5): p. 717-27.
31. Zhang, Z., et al., *Predictive analytics with gradient boosting in clinical medicine.* Ann Transl Med, 2019. **7**(7): p. 152.
32. Chen, T. and C. Guestrin *XGBoost: A Scalable Tree Boosting System*. arXiv e-prints, 2016.
33. Ke, G., et al., *LightGBM: A Highly Efficient Gradient Boosting Decision Tree.* 2017: p. 3146--3154.
34. Prokhorenkova, L., et al. *CatBoost: unbiased boosting with categorical features*. arXiv e-prints, 2017.
35. Podgorelec, V., et al., *Decision trees: an overview and their use in medicine.* J Med Syst, 2002. **26**(5): p. 445-63.
36. Bergstra, J., et al., *Algorithms for hyper-parameter optimization*. Proceedings of the 24th International Conference on Neural Information Processing Systems. 2011, Granada, Spain: Curran Associates Inc. 2546–2554.
37. Guo, C., et al. *On Calibration of Modern Neural Networks*. arXiv e-prints, 2017.
38. *Advances in Large Margin Classifiers*, ed. J.S. Alexander and J.B. Peter. 2000: MIT Press. 412.

39. Zadrozny, B. and C. Elkan, *Transforming classifier scores into accurate multiclass probability estimates*, in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002, ACM: Edmonton, Alberta, Canada. p. 694-699.
40. Kumar, A., P. Liang, and T. Ma *Verified Uncertainty Calibration*. arXiv e-prints, 2019.
41. BRIER, G.W., *VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY.* Monthly Weather Review, 1950. **78**(1): p. 1-3.
42. Niculescu-Mizil, A. and R. Caruana, *Predicting good probabilities with supervised learning*, in *Proceedings of the 22nd international conference on Machine learning*. 2005, ACM: Bonn, Germany. p. 625-632.
43. Lipovetsky, S. and M. Conklin, *Analysis of regression in game theory approach.* Applied Stochastic Models in Business and Industry, 2001. **17**(4): p. 319-330.
44. Štrumbelj, E. and I. Kononenko, *Explaining prediction models and individual predictions with feature contributions.* Knowledge and Information Systems, 2013. **41**: p. 647-665.
45. *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. 1988, Cambridge: Cambridge University Press.
46. Brown, T.B., et al., *Assessment of risk tolerance for adverse events in emergency department chest pain patients: a pilot study.* J Emerg Med, 2010. **39**(2): p. 247-52.
47. Chang, S.H., 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**: p. 6983568.
48. Jobin, A., M. Ienca, and E. Vayena, *The global landscape of AI ethics guidelines.* Nature Machine Intelligence, 2019.
49. Cohen, I.G., et al., *The legal and ethical concerns that arise from using complex predictive analytics in health care.* Health Aff (Millwood), 2014. **33**(7): p. 1139-47.
50. EU, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the*

processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). 2016, Off J Eur Communities. p. 1-88.