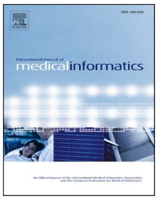
Machine learning for developing a prediction model of hospital admission of emergency department patients: hype or hope?

Hond, A. de; Raven, W.; Schinkelshoek, L.; Gaakeer, M.; Avest, E. ter; Sir, O.; ... ; Groot, B. de

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Machine learning for developing a prediction model of hospital admission of emergency department patients: Hype or hope?

Anne De Hond[a,b,c,*], Wouter Raven[d], Laurens Schinkelshoek[a,b], Menno Gaakeer[e], Ewoud Ter Avest[f], Ozcan Sir[g], Heleen Lameijer[h], Roger Apa Hessels[i], Resi Reijnen[j], Evert De Jonge[k], Ewout Steyerberg[c], Christian H. Nickel[l], Bas De Groot[d]

[a] *Department of Information Technology and Digital Innovation, Leiden University Medical Centre, Albinusdreef 2, 2300 RC, Leiden, the Netherlands*
[b] *Clinical AI Implementation and Research Lab, Leiden University Medical Centre, Albinusdreef 2, 2300 RC, Leiden, the Netherlands*
[c] *Department of Biomedical Data Sciences, Leiden University Medical Centre, Albinusdreef 2, 2300 RC, Leiden, the Netherlands*
[d] *Department of Emergency Medicine, Leiden University Medical Centre, Albinusdreef 2, 2300 RC, Leiden, the Netherlands*
[e] *Department of Emergency Medicine, Adrz Hospital, 's-Gravenpolderseweg 114, 4462 RA, Goes, the Netherlands*
[f] *Department of Emergency Medicine, University Medical Centre Groningen, Hanzeplein1, 9713 GZ, Groningen, the Netherlands*
[g] *Department of Emergency Medicine, Radboud University Medical Centre, Houtlaan 4, 6525 XZ, Nijmegen, the Netherlands*
[h] *Department of Emergency Medicine, Medical Centre Leeuwarden, Henri Dunantweg 2, 8934 AD, Leeuwarden, the Netherlands*
[i] *Department of Emergency Medicine, Elisabeth-TweeSteden Hospital, Doctor Deelenlaan 5, 5042 AD, Tilburg, the Netherlands*
[j] *Department of Emergency Medicine, Haaglanden Medical Centre, Lijnbaan 32, 2512 VA, The Hague, the Netherlands*
[k] *Department of Intensive Care Medicine, Leiden University Medical Centre, Albinusdreef 2, 2300 RC, Leiden, the Netherlands*
[l] *Department of Emergency Medicine, University Hospital Basel, University of Basel, Switzerland*

ARTICLE INFO

Keywords:
Machine learning
Prognosis
Prediction
Emergency medicine
Hospital admission
Triage

ABSTRACT

Objective: Early identification of emergency department (ED) patients who need hospitalization is essential for quality of care and patient safety. We aimed to compare machine learning (ML) models predicting the hospitalization of ED patients and conventional regression techniques at three points in time after ED registration.
Methods: We analyzed consecutive ED patients of three hospitals using the Netherlands Emergency Department Evaluation Database (NEED). We developed prediction models for hospitalization using an increasing number of data available at triage, ~30 min (including vital signs) and ~2 h (including laboratory tests) after ED registration, using ML (random forest, gradient boosted decision trees, deep neural networks) and multivariable logistic regression analysis (including spline transformations for continuous predictors). Demographics, urgency, presenting complaints, disease severity and proxies for comorbidity, and complexity were used as covariates. We compared the performance using the area under the ROC curve in independent validation sets from each hospital.
Results: We included 172,104 ED patients of whom 66,782 (39 %) were hospitalized. The AUC of the multivariable logistic regression model was 0.82 (0.78–0.86) at triage, 0.84 (0.81–0.86) at ~30 min and 0.83 (0.75–0.92) after ~2 h. The best performing ML model over time was the gradient boosted decision trees model with an AUC of 0.84 (0.77–0.88) at triage, 0.86 (0.82–0.89) at ~30 min and 0.86 (0.74–0.93) after ~2 h.
Conclusions: Our study showed that machine learning models had an excellent but similar predictive performance as the logistic regression model for predicting hospital admission. In comparison to the 30-min model, the 2-h model did not show a performance improvement. After further validation, these prediction models could support management decisions by real-time feedback to medical personal.

Abbreviations: ED, emergency department; LOS, length of stay; ML, machine learning.

[*] Corresponding author at: Department of Information Technology and Digital Innovation, Leiden University Medical Centre, Albinusdreef 2, 2300 RC, Leiden, the Netherlands.

E-mail address: a.a.h.de_hond@lumc.nl (A. De Hond).

https://doi.org/10.1016/j.ijmedinf.2021.104496
Received 11 January 2021; Received in revised form 26 April 2021; Accepted 13 May 2021
Available online 15 May 2021

1. Introduction

1.1. Background

Emergency department (ED) crowding is a well-known problem affecting the quality of care and patient safety, also in the Netherlands [1,2]. Long ED length of stay (LOS) is associated with reduced patient satisfaction, negative effects on staff, and poorer patient outcomes, including increased in-hospital mortality [3–6]. ED patients who ultimately need to be admitted contribute disproportionately to the occurrence of crowding [7,8].

1.2. Importance

Reduction of ED-LOS by early identification of patients who need hospitalization has several advantages. First, the hospitalization process can be initialized in parallel to ED management, which would save time and enables fast admission to an appropriate level of care. This has been suggested to reduce mortality [9]. Secondly, patients can anticipate hospitalization, which could increase patient satisfaction. Finally, it may have prognostic value as patients who need hospitalization are often the sickest and will benefit most from time-sensitive ED treatment, i.e., fluid resuscitation in sepsis [8,10].

Unfortunately, the clinical judgment of triage nurses is not good enough to accurately predict the hospitalization of ED patients [11]. ED physicians may produce better risk estimates, but it is uncommon for them to perform triage [12]. Therefore, various regression models have been developed to aid the decision to hospitalize the patient, often with mediocre results [13–18].

The advent of machine learning (ML) and the growing availability of increasingly large databases such as electronic health records offer new opportunities to develop novel prediction models that have a better predictive performance [19–21].

However, recent articles [22,23] state that, on average, the performance of ML was no different from that of logistic regression. Furthermore, a prediction model can only reduce ED-LOS when it has good predictive performance with data available soon after triage. However, some potentially important prognostic patient information (such as vital signs and blood tests) is not available at time of triage. Waiting longer for this information to become available means the ED-LOS reduction will be lower than when deploying soon after triage.

1.3. Aims of this investigation

The aim of the present study was twofold. First, we investigated whether ML models could predict hospitalization of ED patients more accurately than logistic regression. Second, we investigated the trade-off between the potential to improve the predictive performance of the models when including more variables and the potential to reduce time to decision-making by developing models at triage, at ~30 min (when vital signs are available) and ~2 h (when blood test results are available).

2. Methods

2.1. Study design and setting

We used observational multi-center data from the Netherlands Emergency Department Evaluation Database (NEED, for more information, see www.stichting-need.nl), the national quality registry of EDs in the Netherlands. For the present study, data were available of 3 EDs, one tertiary care center, and two urban teaching hospitals. We used data collected between 1 January 2017 and 31 December 2019. The study was approved by the medical ethics committee of the LUMC and registered in the Netherlands Trial Register (NL8743).

2.2. Selection of participants

All consecutive ED patients with a registered presenting complaint in the NEED registry database were prospectively included in the study unless they objected to participating in the registry. We filtered patients at three consecutive time points at which, on average, an increasing number of data become available in the electronic hospital information systems: at triage (~15 min after ED registration), after ~30 min (including all vital signs if measured) and after ~2 h (also including laboratory testing, if performed). For the 15-minute dataset, we excluded patients sent home or referred to a GP within the first 15 min of arrival. It should be kept in mind that these points in time are theoretical and merely indicate the approximate moment when additional data are available in clinical practice, i.e., in the Netherlands, it will take approximately two hours before diagnostic test results are available.

2.3. Data collection

For model development, we used the variables of the Minimal Data Set (MDS) collected in the NEED. For data definitions in the MDS, see Appendix A.

2.4. Variables

2.4.1. Dependent variable

Hospital admission was defined as admission to a normal ward, admission to a medium care or coronary care unit, transfer to another hospital, admission to an intensive care unit, and the patient dying in the ED. The remaining cases were categorized as the patient being discharged. The treating physician was in charge of the decision to hospitalize. Generally, the decision to admit a patient was made after the consultation results and laboratory/radiology testing had become available.

2.4.2. Independent variables

A set of independent variables was identified to predict hospital admission based on a review of the literature [13] and consensus between two ED physicians obtained over multiple discussions involving two ED physicians and two data scientists. The selection was made based on expected relevance and availability. The following variables were considered, depending on the sequential dataset collected (~15 min, ~30 min, and ~2 h after arrival).

Demographics based on age and gender (all models).

Urgency based on referral type, mode of transport, and triage category (all models). The included hospitals used the Manchester Triage System [22] and the similar Netherlands Triage System [23] (both validated tools).

Time of day of presentation (all models).

Presenting complaints categorized in 18 main categories (all models). Presenting complaints of the MTS and NTS systems were merged to form one coherent list (see Appendix B.1.).

Treating specialty of the physician who first saw the patient or to whom the general practitioner referred the patient (all models).

Disease severity based on a continuous (ordinal) Glasgow Coma Scale (all models), vital signs (categorical for the 15-minute model as the outcomes were not available yet at this time point, continuous for the subsequent models), Numeric Rating pain score (NRS; 30-min and 2-h models) and a categorical variable for intravenous fluids administered (2-h model).

Proxies for comorbidity and complexity based on binary indicator variables for blood tests requested, blood cultures, blood gas analysis, radiology imaging, and electrocardiogram (30-min and 2-h models) and a categorical variable for the number of consultations (2-h model) [8].

Laboratory test results (2-h model). We also included whether lab tests were completed for a patient via binary indicator variable (see *Proxies for comorbidity and complexity*) as this signals a certain degree of disease severity.

2.5. Descriptive statistics and model development

The patient population was described with descriptive statistics at each moment after arrival (triage, ~30 min and ~2 h after arrival). Subsequently, we developed four models for each of these moments.

First, we developed a classical statistical multivariable logistic regression model with restricted cubic spline transformations and penalization. It is inherently interpretable: the model equation can be easily written down and understood [24]. However, logistic regression will underperform compared to ML when faced with (highly) complex data patterns.

We also developed two tree-based models: a random forest and a gradient boosted decision trees (XGBoost) model [25]. They perform well in practice, are robust to outliers, and can capture complex relationships. However, they perform poorly on large amounts of categorical data.

Lastly, a deep neural network was developed. This modelling technique has shown exceptional performance in some instances. However, deep neural networks require large amounts of data and have a particular risk of overfitting when using elaborate architectures with respect to sample size.

2.6. Handling of missing data

All values which were unrealistic according to the expert opinion of two ED physicians were set to missing. We removed the observations for which ED location, age, gender, triage category, presenting complaint, and ED length of stay were missing as these were considered crucial in the modeling. For the remaining categorical variables, missing values were assigned a separate category. We imputed the missing value for continuous variables via multiple imputation, and a dummy variable was constructed for each continuous variable indicating where the missing values occurred. The categorical variables were converted into dummy variables, and the continuous variables were normalized.

2.7. Training procedure

We split the data in a train (2/3 of the data) and test dataset (1/3 of the data) stratified by ED location and hospital admission. The train data were used to predict the hospital admission with the abovementioned independent variables. We performed internal-external validation [26]. This is a ‘leave one group out’ cross-validation (where each ED location forms one group) to address the heterogeneity between ED locations throughout the Netherlands [27,28]. We tuned the hyperparameters for the training data on the cross-validated. Subsequently, all models were trained on the entire train dataset with the tuned hyperparameters to arrive at the final models.

2.8. Testing procedure

We applied the models that resulted from the training procedure (2.7) to each ED location separately in the remaining 1/3 of test data. The discriminative performance was measured through the area under the receiver operating characteristic curves (confidence intervals were obtained through bootstrapping). We assessed the calibration through the calibration slope. The test results for the three ED locations were pooled through a random-effects meta-analysis. Sensitivity and specificity were calculated using the cutoff that maximized the sum of sensitivity and specificity. Feature importance was obtained via SHapley Additive exPlanations.

To assess the potential clinical value of these models, we calculated the Mean theoretical reduction in time to decision making based on the thresholds corresponding to the 95 % positive and negative predictive value. A 5% error rate was considered reasonable given the consequences of such an error. Patients retrospectively received an actionable decision (hospitalized or sent home) by the best performing model if their probability of hospitalization was either i) higher than the threshold corresponding to the 95 % PPV or ii) lower than the threshold corresponding to the 95 % NPV. For this set of patients, the time to decision making was adjusted to the model’s time point (15 min, 30 min, or 2 h), and the Mean difference in observed and expected time to decision making was calculated for all patients.

2.9. Software

Descriptive statistics were obtained with IBM SPSS version 25. The main analyses were performed in Python 3.8.0. with R 3.6.3 plug-ins to perform the logistic regression and obtain the pooled results. The code to obtain the results can be obtained upon request.

3. Results

The total number of patients present at the ED decreased over time (Fig. 1 and Table 1). Compared to triage, patients still at the ED after 2 h were on average older, more likely to have arrived by ambulance, had a higher triage category, and were more likely to be admitted to the hospital (Table 1).

After cross-validation (Appendix B.3.), the trained models were validated on the test data. The AUC score (Table 2) of the best performing ML model (XGBoost with AUC 0.84 (0.77−0.88) at triage, 0.86 (0.82−0.89) at ~30 min and 0.86 (0.74−0.93) at ~2 h after arrival) was by and large comparable to that of the logistic regression model (0.82 (0.78−0.86) at triage, 0.84 (0.81−0.86) at ~30 min and 0.83 (0.74−0.90) at ~2 h after arrival). The calibration of all models was generally excellent (Table 2), with calibration slopes close to 1. The XGBoost model had an average sensitivity and specificity of 0.78 and 0.72 at triage, 0.80 and 0.73 at ~30 min, and 0.76 and 0.77 after ~2 h. The models showed minor improvements for the consecutive time points (Table 2). Age and treating specialty were important predictors across all time points (Appendix B.9.-B.11.).

More patients received a decision to be discharged home compared to hospitalization for the 15-minute and 30-min time points (Table 3). For the model at triage, a Mean theoretical time to decision-making reduction of 33 min (25 %) could be realized based on both thresholds across the whole population. At the 30-min time point, this increased to 40 min (26 %), which fell back to 31 min (12 %) at the 2-h point.

4. Limitations

This study has some limitations. All ED locations were used in the training and testing of the models to develop highly generalizable models. An advantage of this approach is that it acknowledges the heterogeneity between locations [27,28]. However, the quest for generalizability might negatively impact the performance at each specific location.

Secondly, the clinician’s decision regarding patient admission was used as the dependent variable for model training. However, the clinical decision-making in itself may be inaccurate, introducing a ceiling effect in terms of the ultimately attainable accuracy of predictive algorithms [29]. Also, patients’ preferences regarding hospitalization or social circumstances might play a role. However, the ceiling effect and effect of patient preferences will be similar for the conventional regression and machine learning models, and therefore the main conclusions remain unchanged.

Finally, consistent with the nature of quality registries, the NEED only contains variables that are registered in the hospital information system. Therefore, vital signs and blood tests were only available for those patients in whom it was measured. Nevertheless, the clinical decision to measure these values contains important prognostic information.

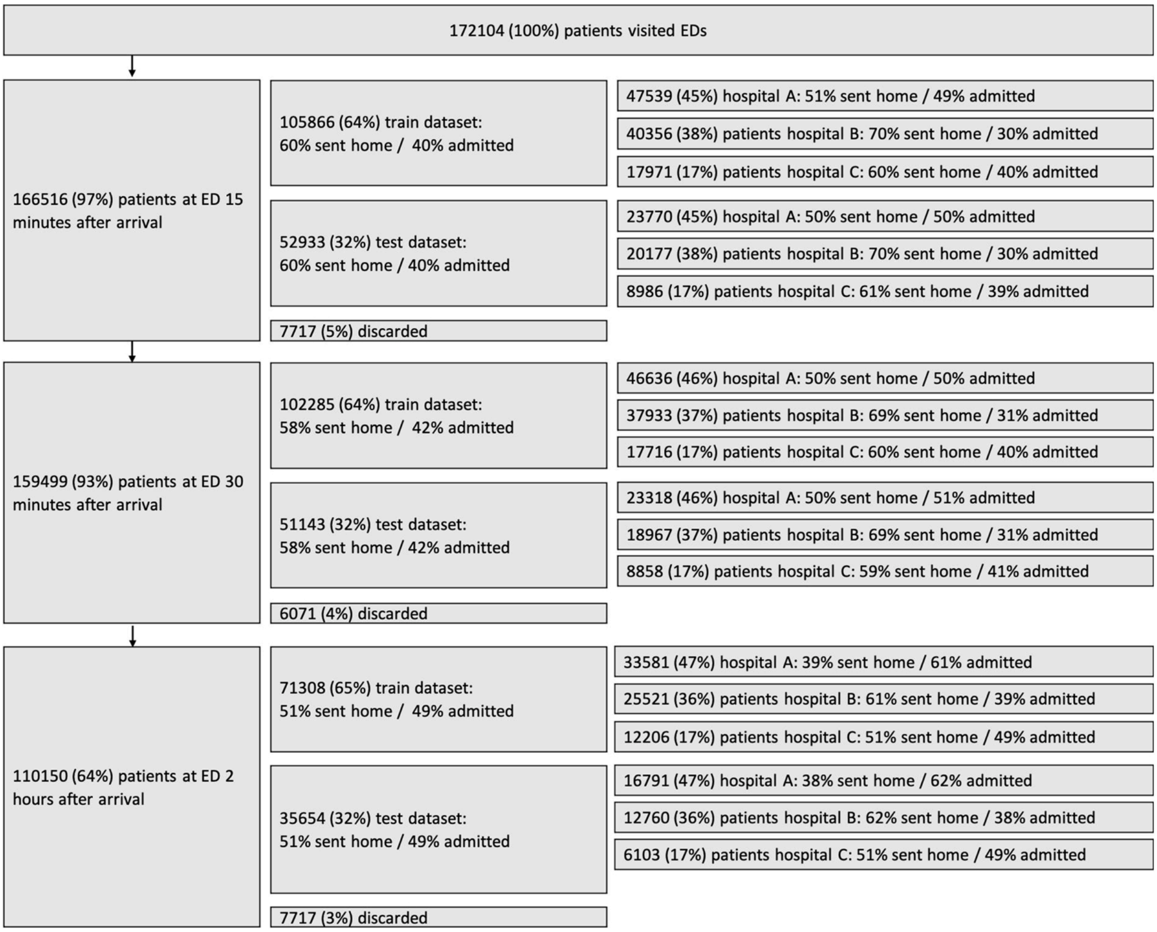

172104 (100%) patients visited EDs
166516 (97%) patients at ED 15 minutes after arrival
105866 (64%) train dataset: 60% sent home / 40% admitted
47539 (45%) hospital A: 51% sent home / 49% admitted
40356 (38%) patients hospital B: 70% sent home / 30% admitted
17971 (17%) patients hospital C: 60% sent home / 40% admitted
52933 (32%) test dataset: 60% sent home / 40% admitted
23770 (45%) hospital A: 50% sent home / 50% admitted
20177 (38%) patients hospital B: 70% sent home / 30% admitted
8986 (17%) patients hospital C: 61% sent home / 39% admitted
7717 (5%) discarded
159499 (93%) patients at ED 30 minutes after arrival
102285 (64%) train dataset: 58% sent home / 42% admitted
46636 (46%) hospital A: 50% sent home / 50% admitted
37933 (37%) patients hospital B: 69% sent home / 31% admitted
17716 (17%) patients hospital C: 60% sent home / 40% admitted
51143 (32%) test dataset: 58% sent home / 42% admitted
23318 (46%) hospital A: 50% sent home / 51% admitted
18967 (37%) patients hospital B: 69% sent home / 31% admitted
8858 (17%) patients hospital C: 59% sent home / 41% admitted
6071 (4%) discarded
110150 (64%) patients at ED 2 hours after arrival
71308 (65%) train dataset: 51% sent home / 49% admitted
33581 (47%) hospital A: 39% sent home / 61% admitted
25521 (36%) patients hospital B: 61% sent home / 39% admitted
12206 (17%) patients hospital C: 51% sent home / 49% admitted
35654 (32%) test dataset: 51% sent home / 49% admitted
16791 (47%) hospital A: 38% sent home / 62% admitted
12760 (36%) patients hospital B: 62% sent home / 38% admitted
6103 (17%) patients hospital C: 51% sent home / 49% admitted
7717 (3%) discarded

Fig. 1. Flow chart of patients at the ED after ~15 min, ~30 min, and ~2 h after arrival at three different locations.
Abbreviations: ED = emergency department.

5. Discussion

5.1. Discussion

Our study showed that machine learning models had an excellent but similar predictive performance as the logistic regression model for predicting hospital admission. Compared to the 30-min model, the 2-h model (including laboratory test results) did not improve performance.

The predictive performance of our models is comparable to other ML and logistic regression models reported in recent literature ([18](N = 506,486) [30];(N = 85,526) [31];(N = 1160) [32];(N = 47,200)) and confirm that – in the current setting – ML models and logistic regression are comparable in performance [18,30–32] with small advantages of modern algorithms. Two of these studies [18,32] also used multi-center data. However, neither one incorporated the potential heterogeneity of the different centers in their training and testing designs, meaning that the general discriminatory performance could be an overestimation of the performance at the individual sites. Also, Peck et al. [31] only included 1160 patients, which might have resulted in a reduction of the predictive power of machine learning models in their study.

A recent study by Barak-Corren, Israelit, and Reis [30] found that laboratory results in a 1 -h model did improve discriminatory performance, in contrast to the findings reported here. This difference with our results may well be explained by the fact that 89 % of patients who had full blood work were hospitalized in the study by Barak-Corren and colleagues. In the NEED, the decision for admission is made after lab results become available.

In only one study [31] did the authors compare their model to the clinical judgment of triage nurses. They found better calibration for the predictions of the models than those of the nurses. We did not directly compare the predictive performance of our models with clinical judgment. However, compared to the pooled sensitivity and specificity of clinical judgment of triage nurses in a recent systematic review [11], our models had slightly higher sensitivity but lower specificity, making their performance roughly comparable.

The present study has several consequences. First, it implies that ML has little benefit for predicting hospital admission over conventional models, at least in the ED setting. ML algorithms may only outperform conventional models if millions rather than hundreds of thousands of patients are included since ML may benefit from a growing sample size [33]. Moreover, the current dataset may lack the covariate complexity that would require the high modeling flexibility ML has to offer. Increasing the number of covariates or the addition of unstructured data could bring to light an advantage of ML over conventional regression methods [18,34].

Although the ML and conventional prediction models had a predictive performance comparable to clinical judgment, they have the advantage that they can be fully automated, and the probability of hospitalization may be reported in the hospital information system, increasing awareness among treating physicians and serving as verification of clinical judgment. Also, as mentioned in the limitations section, it remains to be seen whether clinical judgment should be regarded as the gold standard.

Secondly, although laboratory test results are needed for other purposes such as diagnosis, they appear to have little value for predicting hospital admission in our study. Lab test completion (available after

Table 1
Characteristics split up by time of model.

	Total cohort	Patients, 15 min after arrival	Patients, 30 min after arrival	Patients, 2 h after arrival
Demographics				
N(%)	172,104 (100)	166,516 (100)	159,499 (100)	110,150 (100)
Age, Mean (SD)	49.9(25.2)	50.4 (25.1)	50.9 (25.1)	55.1(23.7)
Gender (female), N (%)	82,812(48.1)	80,544 (48.4)	77,476 (48.6)	54,970(49.9)
Urgency				
Referral type, N(%)				
Self-referral	68,135(39.6)	63,341 (38.0)	58,251 (36.5)	39,579(35.9)
Referral from GP	74,302(43.2)	73,769 (44.3)	72,676 (45.6)	52,742(47.9)
Referral from specialist	27,207(15.8)	26,970 (16.2)	26,171 (16.4)	16,202(14.7)
Missing	2460(1.4)	2436(1.5)	2401(1.5)	1627(1.5)
Arrival by ambulance, N(%)	47,581(27.6)	47,159 (28.3)	46,672 (29.3)	36,975(33.6)
Missing	13,149(7.6)	12,929 (7.8)	12,589 (7.9)	9209(8.4)
Triage category, N (%)				
Blue & green	53,815(31.3)	51,348 (30.8)	47,876 (30.0)	27,014(24.5)
Yellow	68,445(39.8)	67,542 (40.6)	66,053 (41.4)	48,909(44.4)
Orange	36,128(21.0)	36,008 (21.6)	35,600 (22.3)	28,313(25.7)
Red	6216(3.6)	6204(3.7)	6144(3.9)	4251(3.9)
Missing	7500(4.4)	5414(3.3)	3826(2.4)	1663(1.5)
Time of day of presentation 'hh: mm', N(%)				
'00:00–5:59'	13,933(8.1)	13,566 (8.1)	13,148 (8.2)	7943(7.2)
'6:00–11:59'	41,351(24.0)	40,256 (24.2)	38,683 (24.3)	27,188(24.7)
'12:00–17:59'	73,586(42.8)	71,380 (42.9)	68,425 (42.9)	49,121(44.6)
'18:00–23:59'	43,236(25.1)	41,314 (24.8)	39,243 (24.6)	25,898(23.5)
Top 5 Presenting complaints, N(%)				
Extremity problems	36,614(21.3)	35,616 (21.4)	34,067 (21.4)	16,246(14.7)
'Feeling unwell'	26,653(15.5)	26,328 (15.8)	25,740 (16.1)	21,324(19.4)
Abdominal pain	17,425(10.1)	17,248 (10.4)	17,025 (10.7)	14,273(13.0)
Dyspnea	14,369(8.3)	14,296 (8.6)	14,195 (8.9)	12,233(11.1)
Chest pain	12,196(7.1)	12,099 (7.3)	11,897 (7.5)	9399(8.5)
Disease Severity				
Vital score*, N(%)				
Not measured	62,430(36.3)	57,754 (34.7)	52,102 (32.7)	24,100(21.9)
1–4 vital signs measured	58,193(33.8)	57,310 (34.4)	56,100 (35.1)	42,247(38.3)
All vital signs measured	51,481(29.9)	51,452 (30.9)	51,297 (32.2)	43,803(39.8)
GCS, N(%)				
GCS = 15	9745(5.7)	9417(5.7)	9381(5.9)	7767(7.1)
GCS < 15	1385(0.8)	1237(0.7)	1233(0.8)	1005(0.9)
Not assessed	160,974 (93.5)	155,862 (93.6)	148,885 (93.3)	101,378 (92.0)
Pain score, scale 1–10, N(%)				
Not measured	112,030 (65.1)	108,974 (65.4)	104,832 (65.7)	7823(67.0)
1–3	26,277(15.3)	24,927 (15.0)	23,398 (14.7)	14,502(13.2)
4–6	22,672(13.2)	21,796 (13.1)	20,820 (13.1)	14,049(12.8)
7+	11,125(6.5)	10,819 (6.5)	10,449 (6.6)	7776(7.1)
Fluids administered, N(%)				
< 500 mL	11,539(6.7)			9793(8.9)
> 500 mL	12,870(7.5)			11,103(10.1)
None	147,695 (85.8)			89,254(81.0)
Proxies for comorbidity and complexity				
Treating specialty				
Emergency Medicine	33,908(19.7)	33,832 (20.3)	33,182 (21.7)	20,988(19.1)
*Surgery***	35,561(20.7)	35,440 (21.3)	89,118 (55.9)	20,778(18.9)
*Medicine****	90,456(52.6)	90,144 (54.1)	34,683 (21.7)	67,882(61.6)
Missing	12,179(7.1)	7100(4.3)	2516(1.6)	502(0.5)
Number of consultations, N (%)				
None	139,555 (81.1)			89,807(81.5)
One consultation	19,087(11.1)			16,214(14.7)
Two or more consultations	4212(2.4)			3870(3.5)
Missing	9250(5.4)			259(0.2)
Blood tests, N(%)	97,584(56.7)		97,297 (61.0)	82,867(75.2)
Blood cultures, N(%)	13,680(7.9)		13,672 (8.6)	12,761(11.6)
Blood gas analysis, N (%)	22,833(13.3)		22,798 (14.3)	19,831(18.0)
Radiology imaging****, N(%)	94,258(54.8)		92,579 (58.0)	70,736(64.2)
Electrocardiogram, N (%)	43,014(25.0)		42,953 (26.9)	36,845(33.4)
Laboratory tests				
Haemoglobin (mmol/L), median (IQR)[N]	8.4(7.6–9.1) [95,238]			8.4(7.5–9.1) [81,097]
Hematocrit (L/L), median (IQR)[N]	0.40 (0.37–0.43) [94,333]			0.40 (0.37–0.44) [80,245]
Sodium (mmol/L), median (IQR)[N]	140 (137–142) [94,144]			140 (137–141) [80,334]
Leukocytes (x10^9 mg/L), median (IQR)[N]	9.1 (7.0–12.1) [94,067]			9.2 (7.0–12.2) [80,488]
Potassium (mmol/L), median (IQR)[N]	4.1(3.8–4.4) [92,333]			4.1(3.8–4.4) [78,755]
Creatinine (μmol/L), median (IQR)[N]	76(63–96) [92,212]			94(63–97) [79,229]
Urea (mmol/L), median (IQR)[N]	5.7(4.3–7.8) [91,288]			5.7(4.3–7.9) [78,361]
Platelets (x10^9 mg/L), median (IQR) [N]	245 (196–303) [88,955]			245 (195–305) [76,198]

(continued on next page)

Table 1 (*continued*)

	Total cohort	Patients, 15 min after arrival	Patients, 30 min after arrival	Patients, 2 h after arrival
ALAT (U/L), median (IQR)[N]	23(17–34) [83,733]			23(17–34) [72,051]
Gamma GT (U/L), median (IQR)[N]	29(18–57) [83,632]			30(18–59) [71,964]
ASAT (U/L), median (IQR)[N]	25(20–34) [81,503]			25(20–34) [71,964]
CRP (mg/L), median (IQR)[N]	10.5 (4.3–47.0) [80,310]			12.0 (5.0–51.0) [69,378]
Alkalic Fosfate (U/L), median (IQR)[N]	82(66–105) [66,200]			83(67–106) [56,496]
LDH (U/L), median (IQR)[N]	209 (180–105) [65,452]			210 (181–252) [56,132]
Mean Cell Volume (fL), median (IQR) [N]	90(86–93) [62,762]			90(86–93) [53,522]
Neutrophilics (x10ˆ9 mg/L), median (IQR)[N]	6.4(4.5–9.3) [46,297]			6.5(4.6–9.5) [39,597]
Calcium (mmol/L), median (IQR)[N]	2.4(2.3–2.4) [44,752]			2.3(2.3–2.4) [39,435]
Creatine Kinase (U/L), median (IQR) [N]	88(57–143) [35,078]			87(56–141) [29,362]
Hemolysis material present, N(%)				
Yes	4749(2.8)			4166(3.8)
Missing	80,476(46.8)			44,589(40.5)

Patient characteristics are presented for the total cohort and three different times used in the prediction models: after ~15 min, ~30 min, and ~2 h of stay in the emergency department. Normally distributed data are presented as Mean (SD), skewed data as median (IQR), and categorical data as number (%).

Abbreviations: N = number, SD = standard deviation, GCS = Glasgow Coma Scale, n/min = breaths/beats per minute, IQR = interquartile range, mmHg = millimeter of mercury, mL = milliliter, U/L = Units per liter, fL = femtoliter, ED = emergency department.

* Vital signs measured involve: Respiratory Rate, O_2 Saturation, Heart Rate, Systolic Blood Pressure, Diastolic Blood Pressure, and Temperature.

** Surgery contains the specialties of general surgery, traumatology, ophthalmology, orthopedics, otorhinolaryngology, thoracic surgery, urology, gynecology, and neurosurgery.

*** Medicine contains the specialties of internal medicine, cardiology, pulmonology, gastroenterology, neurology, pediatrics, and rheumatology.

**** Radiology imaging is positive if either an X-ray, ultrasound or CT- scan was performed.

Table 2
Pooled random effect meta-analysis performance characteristics.

Dataset	Algorithm	Test AUC (95 % CI)	Calibration slope (95 % CI)
Triage	LR	0.82 (0.78, 0.86)	1.14 (0.92, 1.41)
	RF	0.80 (0.72, 0.85)	1.05 (0.95, 1.17)
	XGBoost	0.84 (0.77, 0.88)	1.09 (0.92, 1.29)
	DNN	0.83 (0.77, 0.88)	1.05 (0.89, 1.24)
~ 30 min	LR	0.84 (0.81, 0.86)	1.12 (0.94, 1.34)
	RF	0.86 (0.83, 0.88)	1.03 (0.90, 1.17)
	XGBoost	0.86 (0.82, 0.89)	1.07 (0.94, 1.21)
	DNN	0.86 (0.82, 0.89)	1.13 (1.01, 1.27)
~ 2 h	LR	0.83 (0.74, 0.90)	1.06 (0.92, 1.23)
	RF	0.86 (0.75, 0.92)	0.98 (0.85, 1.14)
	XGBoost	0.86 (0.74, 0.93)	1.03 (0.92, 1.15)
	DNN	0.86 (0.75, 0.93)	1.02 (0.89, 1.17)

AUC and calibration slope were calculated separately for the three centers and pooled through a random effect meta-analysis for each model.

Abbreviations: *LR* Logistic Regression, *RF* Random Forset, *XGBoost* gradient boosted decision trees, *DNN* Deep Neural Network, *AUC* Area Under the Curve.

Table 3
Potential Mean (relative) time to decision making (TDM) reduction based on number of patients in the test data receiving an earlier decision (admitted or sent home) according to best performing model (XGBoost).

	Total number of patients test data	Number of patients with an actionable decision*	Mean TDM reduction in minutes (Mean relative TDM reduction)** for patients with an actionable decision*	Mean TDM reduction in minutes (Mean relative TDM reduction) for all patients**
Triage				
PPV	52,928	1227 (2%)	174 (90 %)	4.04 (2%)
NPV	52,928	15,281 (29 %)	99.34 (79 %)	28.68 (23 %)
PPV & NPV	52,928	16,508 (31 %)	104.91 (79 %)	32.72 (25 %)
30 min				
PPV	51,137	3200 (6%)	182.29 (83 %)	11.41 (5%)
NPV	51,137	15,369 (30 %)	94.46 (68 %)	28.39 (20 %)
PPV & NPV	51,137	18,569 (36 %)	109.60 (71 %)	39.80 (26 %)
2 h				
PPV	35,649	6000 (17 %)	117.28 (44 %)	19.74 (7%)
NPV	35,649	5706 (16 %)	69.13 (31 %)	11.07 (5%)
PPV & NPV	35,649	11,706 (33 %)	93.81 (38 %)	30.80 (12 %)

*A patient receives an actionable decision from the model when:
i) P(hospitalization) > 95 % PPV threshold for PPV scenario.
ii) P(hospitalization) < 95 % NPV threshold for NPV scenario.
iii) P(hospitalization) > 95 % PPV threshold or P(hospitalization) < 95 % NPV threshold for PPV & NPV combined scenario.

**Mean time to decision making (TDM) and Mean relative TDM reduction in minutes are calculated as: Mean(*TDM patient – TDM patient model*) and Mean (100x(*TDM patient – TDM patient model*) / *TDM patient*). *TDM patient model* is set to 15 min (triage model), 30 min (30-min model), or 2 h (2-h model) for patients with an actionable decision. *TDM patient model* is set to *TDM patient* when the patient did not receive an actionable decision.

Abbreviations: ED Emergency Department, TDM Time to Decision Making, PPV Positive Predictive Value, NPV Negative Predictive Value.

~30 min) may already be a good predictor of hospitalization, regardless of the test result. The decrease in sample size and change in the sample composition (retaining the generally more complex patients in the ED while others are discharged or admitted) over time may also affect predictive performance.

Consequently, the hospitalization process in Dutch EDs could be initialized before test results are available. Based on a prospective study by van der Veen et al. [8] in a similar setting, time to decision making, and therefore ED-LOS could theoretically be reduced by approximately 40 min (see also Table 3), as long as exit blocks are not the main determinant of ED-LOS. As soon as hospital admission is indicated, additional testing could be performed in the clinical decision unit. Note that in clinical practice, an earlier decision may not necessarily translate into a shorter ED-LOS. Patients who are discharged may require other medical attention before being sent home.

Nevertheless, a reduction in the time to decision-making may have other benefits, like helping patients anticipate on the hospitalization, which could increase patient satisfaction. Furthermore, because patients who need hospitalization are often the sickest, it may increase awareness of the treating physician, which could be used during ED management. This type of decision support might also aid patient safety, particularly during the evening and night shifts of inexperienced junior doctors when their supervising consultants are often not present.

5.2. Conclusion

Our study showed that machine learning models had an excellent but similar predictive performance as the logistic regression model in predicting admission. In comparison to the 30-min model, the 2-h model did not show a performance improvement. Future studies should investigate whether larger sample sizes or more variables result in a better predictive performance of ML models. Future research should also examine the clinical effectiveness of implementing of our predictive algorithm including an investigation of the type of circumstances in which one might prefer ML models over classical statistical techniques.

Author's contributions

BdG devised and designed the study, wrote the study protocol, contributed to data collection, contributed to the analyses, and edited the manuscript.

AdeH contributed to the study idea and protocol, did the analyses, and wrote the manuscript.

WR contributed to the study protocol, analyses, and writing and editing of the manuscript.

ES and LS contributed to the analyses and edited the manuscript.

CN contributed to the study protocol and edited the manuscript.

MG, OS, RR, HL, EterA, and RH contributed to data collection.

EdeJ, ES, MG, HL, EterA edited the manuscript.

BdG takes full responsibility for the study as a whole.

All authors read and approved the final manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

SUMMARY TABLE

What was known

- Early identification of emergency department (ED) patients who need hospitalization is essential for quality of care and patient safety.
- The advent of machine learning and the growing availability of increasingly large databases such as electronic health records offer new opportunities to develop novel prediction models with better predictive performance.

What this study added to our knowledge

- Machine learning models have excellent but similar predictive performance as logistic regression for predicting hospital admission.
- While it might be tempting to wait for additional information such as blood test results, they do not improve the machine learning prediction of hospital admission.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Julius de Hond for providing writing assistance during the revision of this article.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijmedinf.2021.104496.

References

[1] M.C. Linden, et al., Drukte op Spoedeisende Hulpafdelingen in nederland: ervaringen van verpleegkundig managers, Triage (2014).
[2] M.C. Van Der Linden, et al., Two emergency departments, 6000km apart: differences in patient flow and staff perceptions about crowding, Int. Emerg. Nurs. 35 (2017) 30–36.
[3] C. Morley, et al., Emergency department crowding: a systematic review of causes, consequences and solutions, PLoS One 13 (8) (2018) e0203316.
[4] S.L. Bernstein, et al., The effect of emergency department crowding on clinically oriented outcomes, Acad. Emerg. Med. 16 (1) (2009) 1–10.
[5] A. Guttmann, 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 342 (2011) d2983.
[6] J.M. Pines, J.E. Hollander, Emergency department crowding is associated with poor care for patients with severe pain, Ann. Emerg. Med. 51 (1) (2008) 1–5.
[7] D.M. Fatovich, Y. Nagree, P. Sprivulis, Access block causes emergency department overcrowding and ambulance diversion in Perth, Western Australia, Emerg. Med. J. 22 (5) (2005) 351–354.
[8] D. van der Veen, et al., Independent determinants of prolonged emergency department length of stay in a tertiary care centre: a prospective cohort study, Scand. J. Trauma Resusc. Emerg. Med. 26 (1) (2018) 81.
[9] C.N.L. Groenland, et al., Emergency department to ICU time is associated with hospital mortality: a registry analysis of 14,788 patients from six university hospitals in the Netherlands, Crit. Care Med. 47 (11) (2019) 1564–1571.
[10] P.B. Patel, M.A. Combs, D.R. Vinson, Reduction of admit wait times: the effect of a leadership-based program, Acad. Emerg. Med. 21 (3) (2014) 266–273.
[11] M.A.M. Afnan, et al., Ability of triage nurses to predict, at the time of triage, the eventual disposition of patients attending the emergency department (ED): a systematic literature review and meta-analysis, Emerg. Med. J. (2020).
[12] R. Bingisser, et al., *Physicians' disease severity ratings are Non-Inferior to the emergency severity index*, J. Clin. Med. 9 (3) (2020).
[13] J.A. Lucke, et al., Early prediction of hospital admission for emergency department patients: a comparison between patients younger or older than 70 years, Emerg. Med. J. 35 (1) (2018) 18–27.
[14] N. Kraaijvanger, et al., Development and validation of an admission prediction tool for emergency departments in the Netherlands, Emerg. Med. J. 35 (8) (2018) 464–470.
[15] Y. Sun, et al., Predicting hospital admissions at emergency department triage using routine administrative data, Acad. Emerg. Med. 18 (8) (2011) 844–850.
[16] C.A. Parker, et al., Predicting hospital admission at the emergency department triage: a novel prediction model, Am. J. Emerg. Med. 37 (8) (2019) 1498–1504.
[17] A. Cameron, et al., A simple tool to predict admission at the time of triage, Emerg. Med. J. 32 (3) (2015) 174–179.
[18] W.S. Hong, A.D. Haimovich, R.A. Taylor, Predicting hospital admission at emergency department triage using machine learning, PLoS One 13 (7) (2018) e0201016.
[19] M. Fernandes, et al., Clinical decision support systems for triage in the emergency department using intelligent systems: a review, Artif. Intell. Med. 102 (2020) 101762.
[20] K. Grant, et al., Artificial intelligence in emergency medicine: surmountable barriers with revolutionary potential, Ann. Emerg. Med. 75 (6) (2020) 721–726.
[21] M. Green, et al., Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room, Artif. Intell. Med. 38 (3) (2006) 305–318.
[22] E. Christodoulou, et al., A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models, J. Clin. Epidemiol. 110 (2019) 12–22.
[23] B.Y. Gravesteijn, et al., Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury, J. Clin. Epidemiol. 122 (2020) 95–107.
[24] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (5) (2019) 206–215.
[25] T. Chen, C. Guestrin, in: XGBoost: A Scalable Tree Boosting System, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery: San Francisco, California, USA, 2016, pp. 785–794.
[26] E.W. Steyerberg, et al., Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: an overview and illustration, Stat. Med. 38 (22) (2019) 4290–4309.
[27] E.W. Steyerberg, Validation in prediction research: the waste by data splitting, J. Clin. Epidemiol. 103 (2018) 131–133.
[28] E.W. Steyerberg, F.E. Harrell Jr., Prediction models need appropriate internal, internal-external, and external validation, J. Clin. Epidemiol. 69 (2016) 245–247.
[29] R. Challen, et al., Artificial intelligence, bias and clinical safety, BMJ Qual. Saf. 28 (3) (2019) 231–237.
[30] Y. Barak-Corren, S.H. Israelit, B.Y. Reis, Progressive prediction of hospitalisation in the emergency department: uncovering hidden patterns to improve patient flow, Emerg. Med. J. 34 (5) (2017) 308–314.
[31] J.S. Peck, et al., Predicting emergency department inpatient admissions to improve same-day patient flow, Acad. Emerg. Med. 19 (9) (2012) E1045–E1054.

[32] X. Zhang, et al., Prediction of emergency department hospital admission based on natural language processing and neural networks, Methods Inf. Med. 56 (5) (2017) 377–389.

[33] A. Halevy, P. Norvig, F. Pereira, The unreasonable effectiveness of data, IEEE Intell. Syst. 24 (2) (2009) 8–12.

[34] E. Ford, et al., Extracting information from the text of electronic medical records to improve case detection: a systematic review, J. Am. Med. Inform. Assoc. 23 (5) (2016) 1007–1015.