From code to clinic: theory and practice for artificial intelligence prediction algorithms

Hond, A.A.H. de

PART II

DEVELOPMENT AND VALIDATION: USE CASES

4

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

Anne de Hond, Wouter Raven, Laurens Schinkelshoek, Menno Gaakeer, Ewoud ter Avest, Ozcan Sir, Heleen Lameijer, Roger Apa Hessels, Resi Reijnen, Evert de Jonge, Ewout Steyerberg, Christian Nickel, and Bas de Groot

4.1 ABSTRACT

4.1.1 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.

4.1.2 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 minutes (including vital signs) and ~2 hours (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.

4.1.3 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 minutes and 0.83 (0.75-0.92) after ~2 hours. 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 minutes and 0.86 (0.74-0.93) after ~2 hours.

4.1.4 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-minute model, the 2-hour model did not show a performance improvement. After further validation, these prediction models could support management decisions by real-time feedback to medical personal.

4.2 INTRODUCTION[2]

4.2.1 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].

4.2.2 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

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

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.

4.2.3 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 hours (when blood test results are available).

4.3 METHODS

4.3.1 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).

4.3.2 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 minutes after ED registration), after ~30 minutes (including all vital signs if measured) and after ~2 hours (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 minutes 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.

4.3.3 Data collection

For model development, we used the variables of the Minimal Data Set (MDS) collected in the NEED.

4.3.4 Variables

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.

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 minutes, ~30 minutes, and ~2 hours 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 supplement S1).

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-minute and 2-hour models) and a categorical variable for intravenous fluids administered (2-hour 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-minute and 2-hour models) and a categorical variable for the number of consultations (2-hour model) [8].
Laboratory test results (2-hour 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.

4.3.5 Descriptive statistics and model development

The patient population was described with descriptive statistics at each moment after arrival (triage, ~30 minutes and ~2 hours 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.

4.3.6 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.

4.3.7 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 during cross-validation. Subsequently, all models were trained on the entire train dataset with the tuned hyperparameters to arrive at the final models.

4.3.8 Testing procedure

We applied the models that resulted from the training procedure 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 corre-

sponding 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 minutes, 30 minutes, or 2 hours), and the Mean difference in observed and expected time to decision making was calculated for all patients.

4.3.9 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.

4.4 RESULTS

The total number of patients present at the ED decreased over time (Figure 4.1 and Table 4.1). Compared to triage, patients still at the ED after 2 hours 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 4.1).

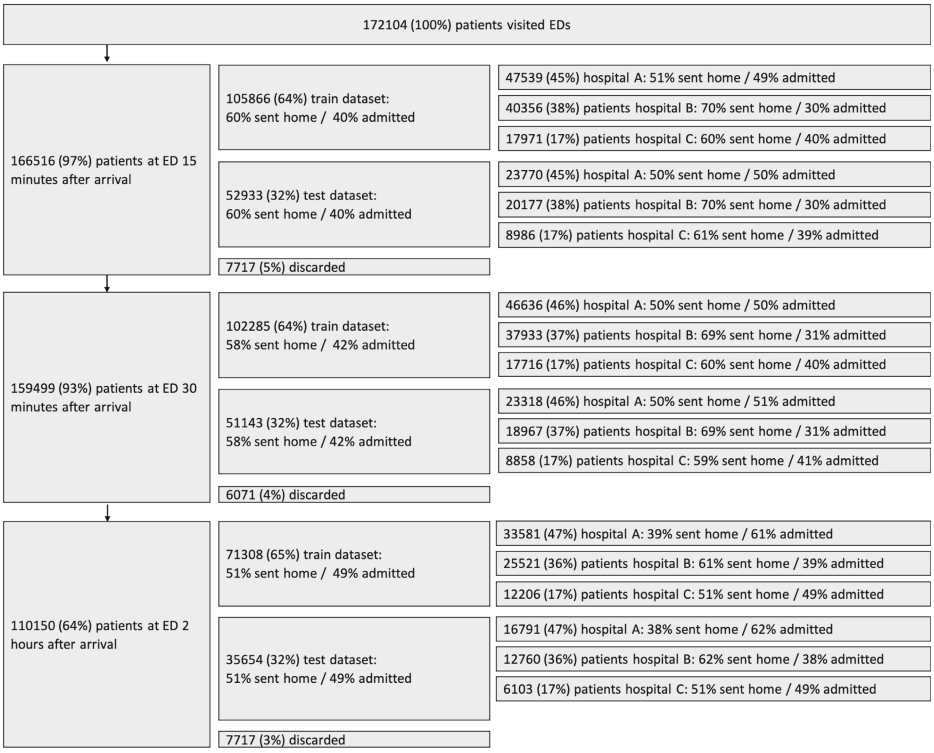


Figure 4.1 Flow chart of patients at the ED after ~15 minutes, ~30 minutes, and ~2 hours after arrival at three different locations

Abbreviations: *ED* emergency department

Table 4.1 Characteristics split up by time of model

	Total cohort	Patients, 15 min after arrival	Patients, 30 min after arrival	Patients, 2 hrs after arrival
Demographics				
N(%)	172104(100)	166516(100)	159499(100)	110150(100)
Age, Mean (SD)	49.9(25.2)	50.4(25.1)	50.9(25.1)	55.1(23.7)
Gender (female), N(%)	82812(48.1)	80544(48.4)	77476(48.6)	54970(49.9)
Urgency				
Referral type, N(%)				
Self-referral	68135(39.6)	63341(38.0)	58251(36.5)	39579(35.9)
Referral from GP	74302(43.2)	73769(44.3)	72676(45.6)	52742(47.9)

Table 4.1 Characteristics split up by time of model (continued)

	Total cohort	Patients, 15 min after arrival	Patients, 30 min after arrival	Patients, 2 hrs after arrival
Referral from specialist	27207(15.8)	26970(16.2)	26171(16.4)	16202(14.7)
Missing	2460(1.4)	2436(1.5)	2401(1.5)	1627(1.5)
Arrival by ambulance, N(%)	47581(27.6)	47159(28.3)	46672(29.3)	36975(33.6)
Missing	13149(7.6)	12929(7.8)	12589(7.9)	9209(8.4)
Triage category, N(%)				
Blue & green	53815(31.3)	51348(30.8)	47876(30.0)	27014(24.5)
Yellow	68445(39.8)	67542(40.6)	66053(41.4)	48909(44.4)
Orange	36128(21.0)	36008(21.6)	35600(22.3)	28313(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'	13933(8.1)	13566(8.1)	13148(8.2)	7943(7.2)
'6:00-11:59'	41351(24.0)	40256(24.2)	38683(24.3)	27188(24.7)
'12:00-17:59'	73586(42.8)	71380(42.9)	68425(42.9)	49121(44.6)
'18:00-23:59'	43236(25.1)	41314(24.8)	39243(24.6)	25898(23.5)
Top 5 Presenting complaints, N(%)				
Extremity problems	36614(21.3)	35616(21.4)	34067(21.4)	16246(14.7)
'Feeling unwell'	26653(15.5)	26328(15.8)	25740(16.1)	21324(19.4)
Abdominal pain	17425(10.1)	17248(10.4)	17025(10.7)	14273(13.0)
Dyspnea	14369(8.3)	14296(8.6)	14195(8.9)	12233(11.1)
Chest pain	12196(7.1)	12099(7.3)	11897(7.5)	9399(8.5)
Disease Severity				
Vital score*, N(%)				
Not measured	62430(36.3)	57754(34.7)	52102(32.7)	24100(21.9)
1-4 vital signs measured	58193(33.8)	57310(34.4)	56100(35.1)	42247(38.3)
All vital signs measured	51481(29.9)	51452(30.9)	51297(32.2)	43803(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	160974(93.5)	155862(93.6)	148885(93.3)	101378(92.0)
Pain score, scale 1 to 10, N(%)				
Not measured	112030(65.1)	108974(65.4)	104832(65.7)	7823(67.0)
1-3	26277(15.3)	24927(15.0)	23398(14.7)	14502(13.2)
4-6	22672(13.2)	21796(13.1)	20820(13.1)	14049(12.8)

Table 4.1 Characteristics split up by time of model (continued)

	Total cohort	Patients, 15 min after arrival	Patients, 30 min after arrival	Patients, 2 hrs after arrival
7+	11125(6.5)	10819(6.5)	10449(6.6)	7776(7.1)
Fluids administered, N(%)				
< 500 ml	11539(6.7)			9793(8.9)
> 500 ml	12870(7.5)			11103(10.1)
None	147695(85.8)			89254(81.0)
Proxies for comorbidity and complexity				
Treating specialty				
Emergency Medicine	33908(19.7)	33832(20.3)	33182(21.7)	20988(19.1)
*Surgery***	35561(20.7)	35440(21.3)	89118(55.9)	20778(18.9)
*Medicine****	90456(52.6)	90144(54.1)	34683(21.7)	67882(61.6)
Missing	12179(7.1)	7100(4.3)	2516(1.6)	502(0.5)
Number of consultations, N(%)				
None	139555(81.1)			89807(81.5)
One consultation	19087(11.1)			16214(14.7)
Two or more consultations	4212(2.4)			3870(3.5)
Missing	9250(5.4)			259(0.2)
Blood tests, N(%)	97584(56.7)		97297(61.0)	82867(75.2)
Blood cultures, N(%)	13680(7.9)		13672(8.6)	12761(11.6)
Blood gas analysis, N(%)	22833(13.3)		22798(14.3)	19831(18.0)
Radiology imaging****, N(%)	94258(54.8)		92579(58.0)	70736(64.2)
Electrocardiogram, N(%)	43014(25.0)		42953(26.9)	36845(33.4)
Laboratory tests				
Haemoglobin (mmol/L), median (IQR)[N]	8.4(7.6-9.1) [95238]			8.4(7.5-9.1) [81097]
Hematocrit (L/L), median (IQR)[N]	0.40(0.37-0.43)[94333]			0.40(0.37-0.44) [80245]
Sodium (mmol/L), median (IQR)[N]	140(137-142) [94144]			140(137-141) [80334]
Leukocytes (x10^9 mg/L), median (IQR)[N]	9.1(7.0-12.1) [94067]			9.2(7.0-12.2) [80488]
Potassium (mmol/L), median (IQR)[N]	4.1(3.8-4.4) [92333]			4.1(3.8-4.4) [78755]
Creatinine (µmol/L), median (IQR)[N]	76(63-96) [92212]			94(63-97) [79229]
Urea (mmol/L), median (IQR)[N]	5.7(4.3-7.8) [91288]			5.7(4.3-7.9) [78361]

Table 4.1 Characteristics split up by time of model (continued)

	Total cohort	Patients, 15 min after arrival	Patients, 30 min after arrival	Patients, 2 hrs after arrival
Platelets (x10^9 mg/L), median (IQR)[N]	245(196-303) [88955]			245(195-305) [76198]
ALAT (U/L), median (IQR)[N]	23(17-34) [83733]			23(17-34)[72051]
Gamma GT (U/L), median (IQR)[N]	29(18-57) [83632]			30(18-59)[71964]
ASAT (U/L), median (IQR)[N]	25(20-34) [81503]			25(20-34) [71964]
CRP (mg/L), median (IQR)[N]	10.5(4.3-47.0)[80310]			12.0(5.0-51.0) [69378]
Alkalic Fosfate (U/L), median (IQR)[N]	82(66-105) [66200]			83(67-106) [56496]
LDH (U/L), median (IQR)[N]	209(180-105) [65452]			210(181-252) [56132]
Mean Cell Volume (fL), median (IQR)[N]	90(86-93) [62762]			90(86-93) [53522]
Neutrophilics (x10^9 mg/L), median (IQR)[N]	6.4(4.5-9.3) [46297]			6.5(4.6-9.5) [39597]
Calcium (mmol/L), median (IQR)[N]	2.4(2.3-2.4) [44752]			2.3(2.3-2.4) [39435]
Creatine Kinase (U/L), median (IQR)[N]	88(57-143) [35078]			87(56-141) [29362]
Hemolysis material present, N(%)				
Yes	4749(2.8)			4166(3.8)
Missing	80476(46.8)			44589(40.5)

Patient characteristics are presented for the total cohort and three different times used in the prediction models: after ~15 minutes, ~30 minutes, and ~2 hours 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, O2 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.

After cross-validation (supplementary Table S2), the trained models were validated on the test data. The AUC score (Table 4.2) 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 minutes and 0.86 (0.74-0.93) at ~2 hours 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 minutes and 0.83 (0.74-0.90) at ~2 hours after arrival). The calibration of all models was generally excellent (Table 4.2), 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 minutes, and 0.76 and 0.77 after ~2 hours. The models showed minor improvements for the consecutive time points (Table 4.2). Age and treating specialty were important predictors across all time points (supplementary Figure S3).

Table 4.2 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 minutes	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 hours	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

More patients received a decision to be discharged home compared to hospitalization for the 15-minute and 30-minute time points (Table 4.3). For the model at triage, a Mean theoretical time to decision-making reduction of 33 minutes (25%) could be realized based on both thresholds across the whole population.

At the 30-minute time point, this increased to 40 minutes (26%), which fell back to 31 minutes (12%) at the 2-hour point.

Table 4.3 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	52928	1227 (2%)	174 (90%)	4.04 (2%)
NPV	52928	15281 (29%)	99.34 (79%)	28.68 (23%)
PPV & NPV	52928	16508 (31%)	104.91 (79%)	32.72 (25%)
30 minutes				
PPV	51137	3200 (6%)	182.29 (83%)	11.41 (5%)
NPV	51137	15369 (30%)	94.46 (68%)	28.39 (20%)
PPV & NPV	51137	18569 (36%)	109.60 (71%)	39.80 (26%)
2 hours				
PPV	35649	6000 (17%)	117.28 (44%)	19.74 (7%)
NPV	35649	5706 (16%)	69.13 (31%)	11.07 (5%)
PPV & NPV	35649	11706 (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 minutes (triage model), 30 minutes (30-minute model), or 2 hours (2-hour 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

4.5 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.

4.6 DISCUSSION

4.6.1 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-minute model, the 2-hour 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-hour 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 ~30 minutes) 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 minutes (see also Table 4.3), 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.

4.6.2 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-minute model, the 2-hour 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.

REFERENCES

1. Linden, M.C., R. Reijnen, R. Derlet, et al., *Drukte op Spoedeisende Hulpafdelingen in Nederland: Ervaringen van verpleegkundig managers.* Triage, 2014.
2. Van Der Linden, M.C., M. Khursheed, K. Hooda, J.M. Pines, and N. Van Der Linden, *Two emergency departments, 6000km apart: Differences in patient flow and staff perceptions about crowding.* Int Emerg Nurs, 2017. **35**: p. 30-36.
3. Morley, C., M. Unwin, G.M. Peterson, J. Stankovich, and L. Kinsman, *Emergency department crowding: A systematic review of causes, consequences and solutions.* PLoS One, 2018. **13**(8): p. e0203316.
4. Bernstein, S.L., D. Aronsky, R. Duseja, et al., *The effect of emergency department crowding on clinically oriented outcomes.* Acad Emerg Med, 2009. **16**(1): p. 1-10.
5. Guttmann, A., M.J. Schull, M.J. Vermeulen, and T.A. Stukel, *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.
6. Pines, J.M. and J.E. Hollander, *Emergency department crowding is associated with poor care for patients with severe pain.* Ann Emerg Med, 2008. **51**(1): p. 1-5.
7. Fatovich, D.M., Y. Nagree, and P. Sprivulis, *Access block causes emergency department overcrowding and ambulance diversion in Perth, Western Australia.* Emerg Med J, 2005. **22**(5): p. 351-4.
8. van der Veen, D., C. Remeijer, A.J. Fogteloo, C. Heringhaus, and B. de Groot, *Independent determinants of prolonged emergency department length of stay in a tertiary care centre: a prospective cohort study.* Scand J Trauma Resusc Emerg Med, 2018. **26**(1): p. 81.
9. Groenland, C.N.L., F. Termorshuizen, W.J.R. Rietdijk, 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, 2019. **47**(11): p. 1564-1571.
10. Patel, P.B., M.A. Combs, and D.R. Vinson, *Reduction of admit wait times: the effect of a leadership-based program.* Acad Emerg Med, 2014. **21**(3): p. 266-73.
11. Afnan, M.A.M., T. Netke, P. Singh, 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. Bingisser, R., S.M. Baerlocher, T. Kuster, R. Nieves Ortega, and C.H. Nickel, *Physicians' Disease Severity Ratings are Non-Inferior to the Emergency Severity Index.* J Clin Med, 2020. **9**(3).
13. Lucke, J.A., J. de Gelder, F. Clarijs, et al., *Early prediction of hospital admission for emergency department patients: a comparison between patients younger or older than 70 years.* Emerg Med J, 2018. **35**(1): p. 18-27.
14. Kraaijvanger, N., D. Rijpsma, L. Roovers, et al., *Development and validation of an admission prediction tool for emergency departments in the Netherlands.* Emerg Med J, 2018. **35**(8): p. 464-470.

15. Sun, Y., B.H. Heng, S.Y. Tay, and E. Seow, *Predicting hospital admissions at emergency department triage using routine administrative data.* Acad Emerg Med, 2011. **18**(8): p. 844-50.
16. Parker, C.A., N. Liu, S.X. Wu, Y. Shen, S.S.W. Lam, and M.E.H. Ong, *Predicting hospital admission at the emergency department triage: A novel prediction model.* Am J Emerg Med, 2019. **37**(8): p. 1498-1504.
17. Cameron, A., K. Rodgers, A. Ireland, R. Jamdar, and G.A. McKay, *A simple tool to predict admission at the time of triage.* Emerg Med J, 2015. **32**(3): p. 174-9.
18. Hong, W.S., A.D. Haimovich, and R.A. Taylor, *Predicting hospital admission at emergency department triage using machine learning.* PLoS One, 2018. **13**(7): p. e0201016.
19. Fernandes, M., S.M. Vieira, F. Leite, C. Palos, S. Finkelstein, and J.M.C. Sousa, *Clinical Decision Support Systems for Triage in the Emergency Department using Intelligent Systems: a Review.* Artificial Intelligence in Medicine, 2020. **102**: p. 101762.
20. Grant, K., A. McParland, S. Mehta, and A.D. Ackery, *Artificial Intelligence in Emergency Medicine: Surmountable Barriers With Revolutionary Potential.* Annals of Emergency Medicine, 2020. **75**(6): p. 721-726.
21. Green, M., J. Björk, J. Forberg, U. Ekelund, L. Edenbrandt, and M. Ohlsson, *Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room.* Artificial Intelligence in Medicine, 2006. **38**(3): p. 305-318.
22. Christodoulou, E., J. Ma, G.S. Collins, E.W. Steyerberg, J.Y. Verbakel, and B. Van Calster, *A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models.* J Clin Epidemiol, 2019. **110**: p. 12-22.
23. Gravesteijn, B.Y., D. Nieboer, A. Ercole, et al., *Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury.* J Clin Epidemiol, 2020. **122**: p. 95-107.
24. Rudin, C., *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.* Nature Machine Intelligence, 2019. **1**(5): p. 206-215.
25. Chen, T. and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, Association for Computing Machinery: San Francisco, California, USA. p. 785–794.
26. Steyerberg, E.W., D. Nieboer, T.P.A. Debray, and H.C. van Houwelingen, *Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration.* Statistics in Medicine, 2019. **38**(22): p. 4290-4309.
27. Steyerberg, E.W., *Validation in prediction research: the waste by data splitting.* J Clin Epidemiol, 2018. **103**: p. 131-133.
28. Steyerberg, E.W. and F.E. Harrell, Jr., *Prediction models need appropriate internal, internal-external, and external validation.* J Clin Epidemiol, 2016. **69**: p. 245-7.
29. Challen, R., J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, *Artificial intelligence, bias and clinical safety.* BMJ Qual Saf, 2019. **28**(3): p. 231-237.
30. Barak-Corren, Y., S.H. Israelit, and B.Y. Reis, *Progressive prediction of hospitalisation in the emergency department: uncovering hidden patterns to improve patient flow.* Emerg Med J, 2017. **34**(5): p. 308-314.

31. Peck, J.S., J.C. Benneyan, D.J. Nightingale, and S.A. Gaehde, *Predicting emergency department inpatient admissions to improve same-day patient flow.* Acad Emerg Med, 2012. **19**(9): p. E1045-54.
32. Zhang, X., J. Kim, R.E. Patzer, S.R. Pitts, A. Patzer, and J.D. Schrager, *Prediction of Emergency Department Hospital Admission Based on Natural Language Processing and Neural Networks.* Methods Inf Med, 2017. **56**(5): p. 377-389.
33. Halevy, A., P. Norvig, and F. Pereira, *The Unreasonable Effectiveness of Data.* IEEE Intelligent Systems, 2009. **24**(2): p. 8-12.
34. Ford, E., J.A. Carroll, H.E. Smith, D. Scott, and J.A. Cassell, *Extracting information from the text of electronic medical records to improve case detection: a systematic review.* J Am Med Inform Assoc, 2016. **23**(5): p. 1007-15.

SUPPLEMENTARY MATERIAL

S1 Synchronization of MTS and NTS presenting complaints

The participating EDs in the study made use of different triage systems for the registration of presenting complaints. Both the Netherlands Triage System (NTS) (included 50 presenting complaints) and the Manchester Triage System (MTS) (included 51 presenting complaints) were used. In order to use both MTS and NTS presenting complaints in the logistic regression and machine learning models, we merged the MTS and NTS presenting complaints into one combined list of 51 complaints as shown below. Whenever presenting complaints present in either the MTS or NTS could not be matched with complaints present in the other triage system, a distinct presenting complaint was made to be used in the study (e.g. Abscesses & local infections).

Table S1 Synchronization of MTS and NTS presenting complaints

Synchronized presenting complaints used in the study (51 in total)	**MTS presenting complaints (51 in total)**	**NTS presenting complaints (50 in total)**
Abdominal pain	Abdominal pain in adults Abdominal pain in children	Abdominal pain in adults Abdominal pain in children
Abscesses & local infections	Abscesses and local infections	
Allergy, bites & stings	Allergy Bites and stings	Allergic reaction and stings
Apparently drunk	Apparently drunk	
Assault	Assault	
Asthma	Asthma	
Back pain	Back pain	Back pain
Behaving strangely or suicidal	Behaving strangely	Behaving strangely or suicidal
Breast infection		Breast infection
Burns & scalds	Burns and scalds	Burns and scalds
Chest pain	Chest pain	Chest pain
Collapse	Collapsed adult	Collapse Dizziness
Constipation		Constipation
Coughing		Coughing
Crying baby	Crying baby	
Dental problems	Dental problems	Dental problems
Diabetes	Diabetes	Diabetes
Diarrhea & vomiting	Diarrhea and vomiting	Diarrhea Vomiting

Table S1 Synchronization of MTS and NTS presenting complaints (continued)

Synchronized presenting complaints used in the study (51 in total)	**MTS presenting complaints (51 in total)**	**NTS presenting complaints (50 in total)**
Dyspnea	Shortness of breath in adults Shortness of breath in children	Shortness of breath
Ear problems	Ear problems	Ear problems
Exposure to chemicals	Exposure to chemicals	
Extremity problems	Limb problems	Arm problems General/limb trauma Leg problems
Eye problems	Eye problems	Eye problems
Facial problems	Facial problems	Facial trauma Nosebleed
Falls	Falls	
Feeling unwell	Unwell adult Unwell child	Fever in adults Fever in children Neurological failure Unwell adult Unwell child
Fits	Fits	Fits
Foreign body	Foreign body	Foreign body
Gastro-intestinal (GI) bleeding	Gastro-intestinal (GI) bleeding	
Genital problems	Testicular pain	Genital problems
Headache	Headache	Headache
Implantable Cardioverter Defibrillator (ICD)		Implantable Cardioverter Defibrillator (ICD)
Irritable child	Irritable child	
Limping child	Limping child	
Major incidents – primary	Major incidents – primary	
Mental illness	Mental illness	
Near-drowning		Near-drowning
Neck pain	Neck pain	Neck problems Neck trauma
Overdose & poisoning	Overdose and poisoning	Poisoning
Palpitations	Palpitations	Palpitations
Per vaginum (VP) bleeding	Per vaginum (VP) bleeding	Per vaginum (VP) bleeding
Pregnancy	Pregnancy	Childbirth
Rashes	Rashes	Rashes

Table S1 Synchronization of MTS and NTS presenting complaints (continued)

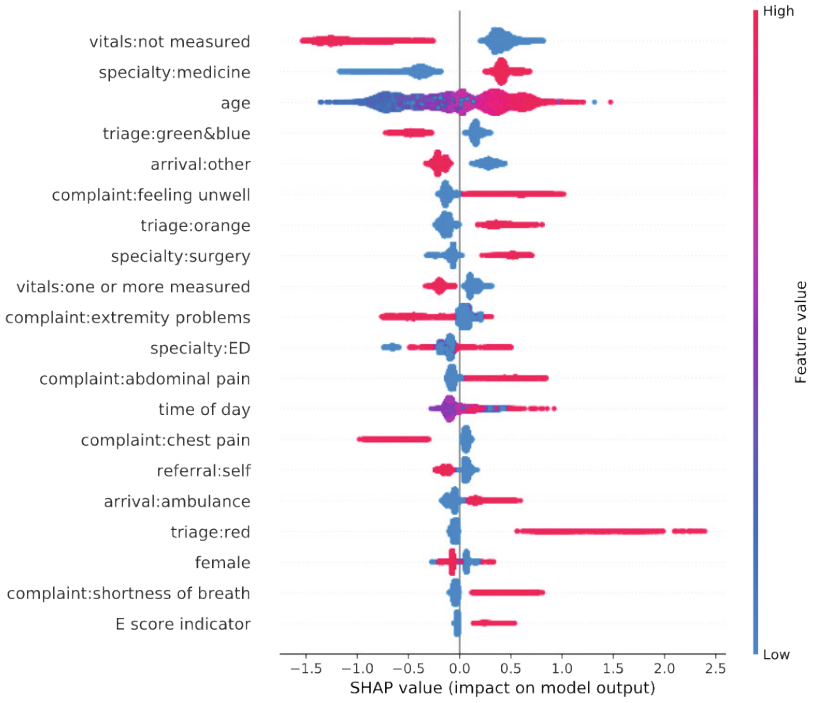
Synchronized presenting complaints used in the study (51 in total)	MTS presenting complaints (51 in total)	NTS presenting complaints (50 in total)
Rectal problems		Rectal problems
Self-harm	Self-harm	
Sexually acquired infection	Sexual acquired infection	
Throat problems	Sore throat	Throat problems
Trauma	Head injury Major trauma Torso injury	Abdominal trauma Back trauma Head trauma Thorax trauma
Urinary problems	Urinary problems	Urinary problems
Worried parent	Worried parent	
Wounds	Wounds	Wounds

Bold presenting complaints directly used in the modelling, remaining complaints grouped in category 'other'.
Abbreviations: *MTS* Manchester Triage System, *NTS* Netherlands Triage System

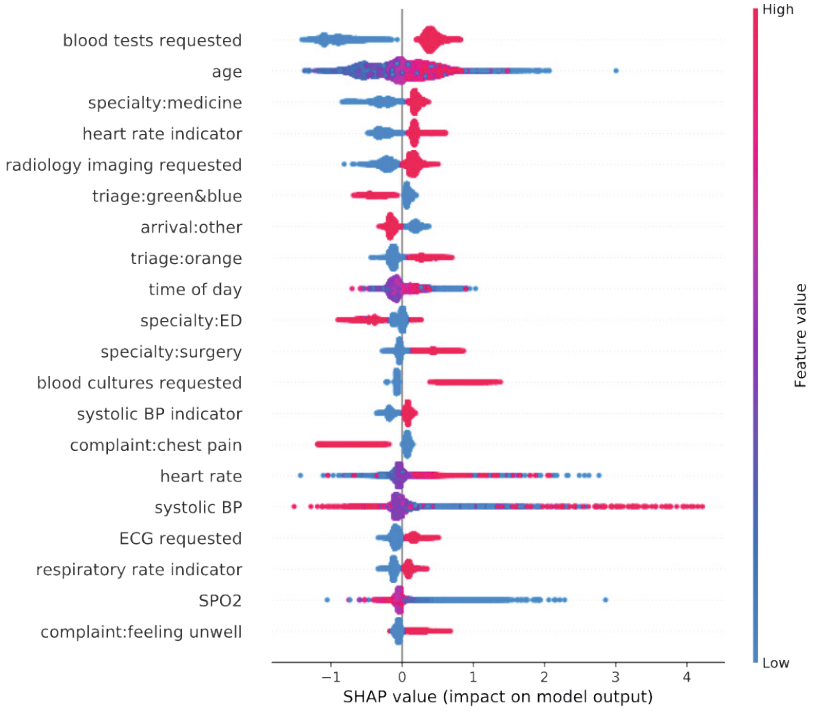
Table S2 Cross-validation resulting hyperparameters and AUC

	Triage	30 minutes	2 hours
LR			
Penalty	20	200	200
AUC	0.81	0.82	0.81
RF			
Estimators	2000	1500	2000
AUC	0.78	0.83	0.82
XGBoost			
Estimators	500	500	500
AUC	0.82	0.84	0.82
DNN			
Nodes	120	120	120
Layers	3	2	3
AUC	0.81	0.84	0.82

a)


vitals:not measured
specialty:medicine
age
triage:green&blue
arrival:other
complaint:feeling unwell
triage:orange
specialty:surgery
vitals:one or more measured
complaint:extremity problems
specialty:ED
complaint:abdominal pain
time of day
complaint:chest pain
referral:self
arrival:ambulance
triage:red
female
complaint:shortness of breath
E score indicator
−1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5
SHAP value (impact on model output)
High
Feature value
Low

b)


blood tests requested
age
specialty:medicine
heart rate indicator
radiology imaging requested
triage:green&blue
arrival:other
triage:orange
time of day
specialty:ED
specialty:surgery
blood cultures requested
systolic BP indicator
complaint:chest pain
heart rate
systolic BP
ECG requested
respiratory rate indicator
SPO2
complaint:feeling unwell
−1 0 1 2 3 4
SHAP value (impact on model output)
High
Feature value
Low

c)

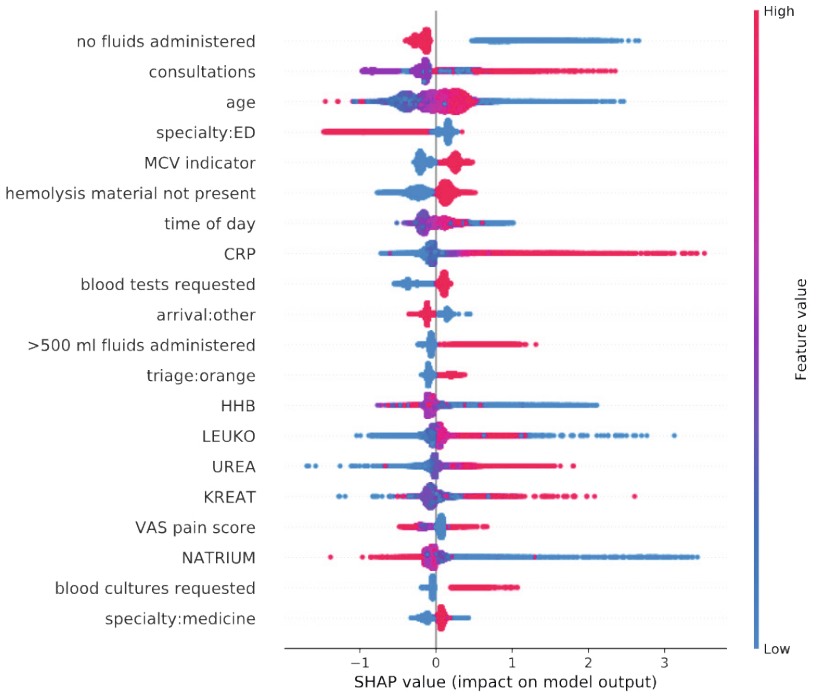


Figure S3 Shapley values for the best performing model (XGBoost) a) at triage, b) at ~30 minutes, and c) at ~2 hours