Available online at www.sciencedirect.com

Resuscitation Plus

journal homepage: www.journals.elsevier.com/resuscitation-plus

Clinical paper

Machine learning model predicts short-term mortality among prehospital patients: A prospective development study from Finland

Joonas Tamminen[a,b,1,*], *Antti Kallonen*[a,1], *Sanna Hoppu*[b], *Jari Kalliomäki*[b,c]

[a] *Faculty of Medicine and Health Technology, Tampere University, PO Box 2000, FI-33521 Tampere, Finland*
[b] *Emergency Medical Services, Tampere University Hospital, PO Box 2000, FI-33521 Tampere, Finland*
[c] *Intensive Care Medicine, Tampere University Hospital, PO Box 2000, FI-33521 Tampere, Finland*

Abstract

Aim: To show whether adding blood glucose to the National Early Warning Score (NEWS) parameters in a machine learning model predicts 30-day mortality more precisely than the standard NEWS in a prehospital setting.

Methods: In this study, vital sign data prospectively collected from 3632 unselected prehospital patients in June 2015 were used to compare the standard NEWS to random forest models for predicting 30-day mortality. The NEWS parameters and blood glucose levels were used to develop the random forest models. Predictive performance on an unknown patient population was estimated with a ten-fold stratified cross-validation method.

Results: All NEWS parameters and blood glucose levels were reported in 2853 (79%) eligible patients. Within 30 days after contact with ambulance staff, 97 (3.4%) of the analysed patients had died. The area under the receiver operating characteristic curve for the 30-day mortality of the evaluated models was 0.682 (95% confidence interval [CI], 0.619−0.744) for the standard NEWS, 0.735 (95% CI, 0.679−0.787) for the random forest-trained NEWS parameters only and 0.758 (95% CI, 0.705−0.807) for the random forest-trained NEWS parameters and blood glucose. The models predicted secondary outcomes similarly, but adding blood glucose into the random forest model slightly improved its performance in predicting short-term mortality.

Conclusions: Among unselected prehospital patients, a machine learning model including blood glucose and NEWS parameters had a fair performance in predicting 30-day mortality.

Keywords: Machine learning, Prehospital, Risk stratification, NEWS

Introduction

Various early warning score (EWS) systems have been introduced to facilitate clinical decision-making in hospital wards; their aim is to detect an inpatient's physiological deterioration prior to adverse outcomes.[1−4] These systems report an aggregate score of physiological measurements of the patient's vital functions. A higher score indicates an increased risk of a short-term medical emergency (e.g. 24-h, 48-h and 30-day mortality, admission to an intensive care unit [ICU] or sepsis).

The signs of impending physiological deterioration and subsequent cardiac arrest can be observed hours before cardiovascular collapse,[5,6] and the Royal College of Physicians advocates the use of the National Early Warning Score (NEWS) also in the prehospital setting.[1] However, the performance of any prehospital EWS system to predict short-term mortality is modest, as only the extreme aggregate scores (i.e. NEWS = 0 or 7) predict a clinically relevant outcome.[7,8]

* Corresponding author at: Medical School, University of Tampere and Emergency Medical Services, Tampere University Hospital, PO Box 2000, FI-33521 Tampere, Finland.
E-mail address: joonas.i.tamminen@tuni.fi (J. Tamminen).
[1] Equal contribution.

http://dx.doi.org/10.1016/j.resplu.2021.100089
Received 27 September 2020; Received in revised form 18 January 2021; Accepted 20 January 2021

Therefore, the standard NEWS' predictive performance should be further strengthened, especially for moderate-risk patients. Retrospective data suggest that adding blood glucose level to NEWS in the prehospital setting and some inflammatory biomarkers to NEWS in the emergency department might improve its performance.[9,10] In addition, modern machine learning methods tailored to a given patient population, such as the random forest (RF) method, seem to outperform traditional logistic regression models in predicting mortality among hospitalised ward patients.[11] RF is a modern machine learning method based on multiple randomly derived decision trees.[12]

We hypothesised that RF algorithms based on readily available physiological measurements would outperform the standard NEWS in the prehospital setting for predicting adverse outcomes. This development study compared the standard NEWS' diagnostic performance to that of RF algorithms trained with NEWS parameters and blood glucose levels for predicting 30-day mortality in unselected adult prehospital patients.

Methods

Design

This descriptive cohort study was conducted in the Tampere University Hospital (Tays) District, Finland. The city of Tampere and the surrounding rural and suburban areas cover a population of 520,000.[13] The emergency medical services (EMS) system comprises first-response units and basic level ambulances, advanced-level ambulances and a physician-staffed helicopter emergency services unit. The study area has one tertiary hospital, one regional hospital and 18 municipal primary health care centres.

The need for informed patient consent was waived, since the study design was observational, involving no interventions to standard therapy. The Tays Ethics Committee reviewed the study protocol (approval no: R10111, May 5th 2015).

Study cohort

The study cohort consisted of all consecutive adult patients (age $\geq$18 years) that the EMS personnel encountered from June 1st 2015 up to and including June 30th 2015. Cases with unknown civil registration numbers or missing case report forms, EMS-encountered cardiac arrest or EMS-confirmed death at the scene, in terminal care, transported to other hospital districts or encountered by EMS units from another district were excluded, since calculation of NEWS would be inappropriate or unfeasible in such cases.

Outcomes

The primary outcome was 30-day mortality. The patient mortality data were retrospectively extracted from the Digital and Population Data Services Agency. The secondary outcomes were 24-h and 48-h mortality, ICU admission and a composite outcome of 48-h mortality or ICU admission.

Predictors

The predictor variables of NEWS and the RF models were prospectively collected, and NEWS scores were retrospectively calculated. During the study period, the EMS was mandated to complete all NEWS parameters (i.e. respiration rate, oxygen saturation [SpO_2], administration of supplemental oxygen, systolic blood pressure, heart rate, level of consciousness and temperature) in every encountered patient regardless of the mission type at the scene before any intervention. The completeness of the NEWS parameters was verified by medical students in the emergency department of the tertiary hospital during the data collection. In the emergency department, there were altogether six medical students who worked in different shifts around the clock. The medical students audited the paper CFRs by re-checking the medical reports. A second audit was made by the author J.K while he transferred the paper CRFs to a digital format.

Contrary to the standard NEWS, the level of consciousness was assessed with the Glasgow Coma Scale, and it was entered as a categorical predictor variable into the RF models. In addition to the standard NEWS parameters, blood glucose level was included as a continuous variable in the RF models. Clinical judgement was used to ascertain whether the patient's blood glucose level was measured. The indications for measuring blood glucose were (1) known type 1 or type 2 diabetes, (2) altered level of consciousness or (3) suspected acute myocardial infarction or stroke. If the same patient had multiple contacts with the EMS personnel during the one-month study period, only the first contact was included in the analysis. Additionally, a sensitivity analysis based on the last contact in the study period was performed.

Sample size and missing data

The study material was collected for a manuscript in preparation which shares the same raw data but has a different aim and design. Since the present study was a post hoc analysis, no formal sample size calculations were performed for this research question. The development of the models was a complete-case analysis in which patients with any missing NEWS parameter or unknown blood glucose level were excluded.

Statistical analysis

All statistical analyses were done using Python language version 3.6.9 or R version 4.0.0. The main statistical packages used were NumPy version 1.17.3 and sklearn version 0.21.3 for Python. Continuous data were presented as means or medians and standard deviations or interquartile ranges, respectively, and categorical data were reported in frequencies and percentiles. The comparison between the groups was performed using a chi-squared test for the categorical data and a Mann–Whitney U-test for the continuous data.

Model development

RF was selected as a machine learning method for this study since it has outperformed logistic regression and the Modified Early Warning Score in in-hospital settings.[10] In our study, two RF models were developed: (1) an RF model derived from NEWS parameters only and (2) an RF model derived from NEWS parameters and blood glucose levels. Since additional input features are not detrimental to the RF model's performance, we decided to use all input features in the model development. The RF models were developed by applying ten-fold stratified cross-validation,[14] where each fold presents an independent subset of

the data to the RF algorithm to train on and uses another subset to estimate predictive performance with an area under the receiver operating characteristics curve (AUROC) performance metric. Stratified division of the folds was used to keep the ratio of deceased patients in the training data the same as in the whole population.

Confidence intervals [CIs] for the cross-validated AUROC scores were calculated using bootstrapping with 10,000 sample points. This bootstrapped distribution of AUROC scores may exhibit non-normal distribution, so the intervals were calculated numerically using the sampled bootstrap distribution to make sure the values were representative.

Model comparisons

Performance of the different RF models and the standard NEWS was compared using the same cross-validation folds for each classifier. To make NEWS scores comparable to a supervised machine learning method, a dummy classifier was designed, which is able to output the score for a cross-validation fold. The bootstrapping method was also used to estimate p-values that were numerically calculated.

Results

EMS was dispatched to 6202 missions, and 4994 prehospital patients were contacted by ambulance personnel during the one-month study period. Of these patients, 3632 met our inclusion criteria. A total of 2853 (79%) patients had complete vital sign data and were included in the primary endpoint analysis (Fig. 1). A minority of the eligible patients with all the NEWS variables measured were excluded due to a missing blood glucose level (96/2949; [3.3%]). All missing vital signs in the eligible patients are presented in Table S1 in the Supplementary Appendix.

The study population's baseline characteristics are presented in Table 1. The 3632 patients eligible for analysis and the 2853 analysed patients were similar in terms of NEWS parameters, blood glucose level, 30-day, 24-h and 48-h mortality and ICU admission. A majority of the study population were low risk patients (i.e. had a NEWS score 0 −4). The mean age of the analysed patients was slightly higher than that of the eligible patients (66 years vs. 63 years, $p < 0.001$). Over one-third of the patients were left at the scene (34% [957] of the analysed patients and 36% [1313] of the eligible patients, $p = 0.34$).

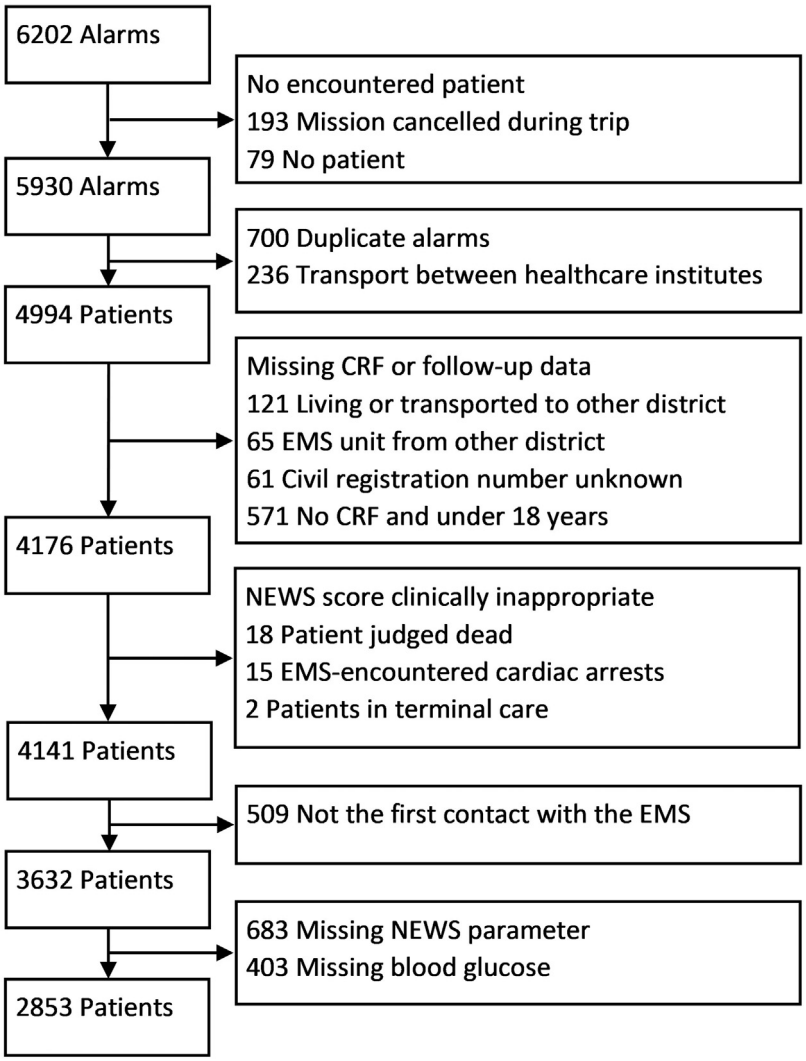

Fig. 1 – Formation of the study population. CRF = case report form; EMS = emergency medical services; NEWS = national early warning score.

Table 1 – Baseline characteristics.

N	Analysed patients 2853	Eligible patients 3632
Age, mean (SD); years	66 (21)	63 (21)
Male sex, %	50	50
NEWS score, median (IQR)	1 (0–3)	1 (0–3)
0, n (%)	735 (26)	1057 (29)
Total 1–4, n (%)	1721 (60)	2122 (58)
3 in single parameter, n (%)	607 (21)	704 (19)
Total 5–6, n (%)	195 (6.8)	228 (6.3)
Total 7 or more, n (%)	202 (7.1)	225 (6.2)
Respiration rate, median (IQR); min^{-1}	16 (15–18)	16 (15–18)
Oxygen saturation, median (IQR); %	97 (95–98)	97 (95–98)
Any supplemental oxygen, %	8.2	7.6
Temperature, median (IQR); °C	36.7 (36.2–37.1)	36.7 (36.3–37.1)
Systolic blood pressure, median (IQR); mmHg	143 (127–164)	143 (127–163)
Heart rate, median (IQR); min^{-1}	85 (72–100)	86 (73–100)
Glasgow Coma Scale >13, %	94	94
Blood glucose, median (IQR); mmol/l	6.7 (5.7–8.2)	6.6 (5.6–8.2)
Glasgow Coma Scale, median (IQR)	15 (15–15)	15 (15–15)
Transportation to, %		
Emergency department	40	38
General practitioner	19	19
Central hospital	6	5
Detoxification centre or jail	2	2
Not transported	34	36
30-day mortality, n (%)	97 (3.4)	114 (3.1)
24-h mortality, n (%)	13 (0.5)	16 (0.4)
48-h mortality, n (%)	18 (0.6)	22 (0.6)
ICU admission, n (%)	32 (1.1)	46 (1.3)
ICU admission/48-h mortality, n (%)	49 (1.7)	66 (1.8)

SD = standard deviation; NEWS = National Early Warning Score; IQR = interquartile range; ICU = intensive care unit.

Within a month after contact with the EMS personnel, 114 (3.1%) eligible patients died. Of these patients, 97 (84%) were analysed.

Fig. 2 shows the receiver operating characteristic (ROC) curves for 30-day mortality. The ROC curves for the secondary outcomes are presented in the supplementary appendix (Fig. S3). Table 2 summarises the cross-validated AUROCs with bootstrapped CIs and p-values for pairwise comparison. The RF models had greater AUROC for 30-day mortality than NEWS (NEWS 0.682 [95% CI, 0.619–0.744]; RF including NEWS parameters was 0.735 [95% CI, 0.679–0.787], p = 0.008 compared with NEWS; RF including NEWS parameters and blood glucose was 0.758 [95% CI, 0.705–0.807], p < 0.001 compared with NEWS).

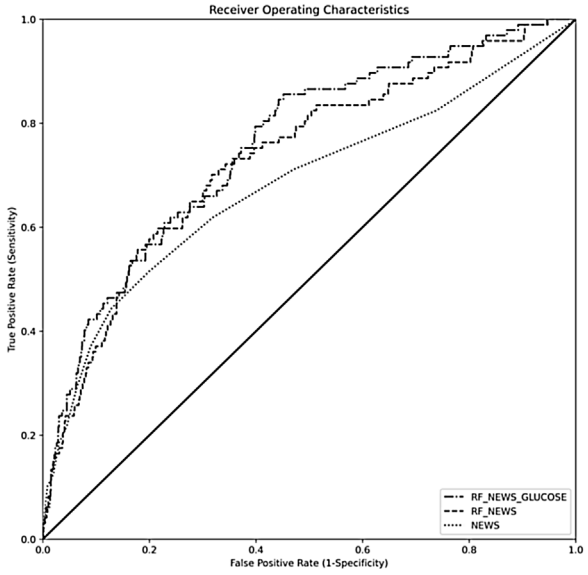

Fig. 2 – Area under the receiver operating characteristics curves for 30-day mortality.

In relation to the secondary outcomes, the AUROCs for the two RF models did not differ from the standard NEWS, but NEWS and both the RF models performed well in predicting 24-h mortality (NEWS 0.895 [95% CI, 0.816–0.961], RF including NEWS parameters was 0.899 [95% CI, 0.811–0.954] and RF including NEWS parameters and blood glucose was 0.953 [95% CI, 0.927–0.976]). The AUROCs for 48-h mortality, ICU admission and their combination were similar in all three models. A sensitivity analysis based on the last contact showed only minor changes to the models' performance (Table S4).

Discussion

Principal findings

In this prospective study, we collected NEWS parameters and blood glucose levels and developed machine learning algorithms to predict 30-day mortality among unselected adult emergency patients. We found that the RF models performed better in predicting 30-day mortality than the standard NEWS. However, the clinical significance of this finding could be questioned as the 95% CIs for the AUROCs are overlapping to a rather large degree. Regarding the secondary

Table 2 – AUROCs with 95% confidence intervals and pairwise comparisons for the cross-validated models.

	NEWS	RF 1	RF 2	p-value		
				NEWS vs RF 1	NEWS vs RF 2	RF 1 vs RF 2
30-d mortality	0.682 (0.619–0.744)	0.735 (0.679–0.787)	0.758 (0.705–0.807)	0.008	<0.001	0.074
24-h mortality	0.890 (0.797–0.966)	0.875 (0.707–0.976)	0.940 (0.860–0.985)	0.89	0.36	0.46
48-h mortality	0.845 (0.729–0.936)	0.808 (0.629–0.957)	0.881 (0.751–0.972)	0.52	0.32	0.12
ICU admission	0.806 (0.715–0.887)	0.807 (0.714–0.890)	0.814 (0.726–0.892)	0.94	0.73	0.72
ICU admission or 48-h mortality	0.818 (0.749–0.882)	0.811 (0.739–0.877)	0.847 (0.785–0.902)	0.74	0.07	0.09

AUROC = area under the receiver operating characteristics curve; NEWS = National Early Warning Score; RF 1 = random forest trained with NEWS parameters only; RF 2 = random forest trained with NEWS parameters and glucose; ICU = intensive care unit.

outcomes, including 48-h mortality and ICU admission, the standard NEWS and the RF model that included blood glucose levels performed equally. That RF model showed excellent performance in predicting 24-h mortality.

Relation of results to other studies

Machine learning models have been developed for various medical purposes.[15] In emergency medicine, speech recognition has been proposed to enhance dispatch, and a machine learning model has been tested for risk stratification at emergency departments.[16,17] Some in-hospital studies have used patients' vital signs and laboratory tests to predict physiological deterioration, sepsis or in-hospital cardiac arrest.[11,18] A recent prehospital study used an RF model to evaluate predictors of 30-day survival in patients with out-of-hospital cardiac arrest.[19]

To the best our knowledge, only one study has also used readily available information in a prehospital setting to train a machine learning model as a risk stratification tool.[20] Spangler et al. found that a different machine learning method to ours (XGBoost) trained with ambulance record data (i.e. vital signs, patient demographics and mission characteristics) was superior to the traditional NEWS in predicting 48-h mortality (AUROC for NEWS, 0.85 [95% CI, 0.83 −0.87] vs. AUROC for XGBoost, 0.89 [95% CI, 0.87−0.91]). Contrary to this study, their study population was more selective, as they excluded patients left at the scene or transported to the non-emergency department. In our material, a third of the patients were left at the scene and a quarter of the patients had a NEWS score of 0, which may indicate that our patient population was less severely ill.

Disturbances in glucose homeostasis might precede impending physiological deterioration or be its consequence in diabetic and non-diabetic emergency patient populations.[9,21] Vihonen et al. found that the standard NEWS and their NEWSgluc logistic regression model had similar AUROCs for 30-day and 24-h mortality, but severe hypoglycaemia was noted to be an important prehospital predictor for death at 30 days (blood glucose 3.0 mmol/L or less; unadjusted odds ratio, 2.06 [95% CI, 1.28−3.19]). However, their study had a notable selection bias, as only 4% of the included patients were analysed. In our study, we observed that measuring blood glucose when clinically appropriate slightly improved the RF model's ability to predict 24-h, 48-h and 30-day mortality. This may indicate that an elevated blood glucose level and stress hyperglycaemia should be suspected at a low threshold among moderate-risk emergency patients but not be measured routinely.

Clinical implications

Machine learning algorithms could be utilised more extensively in the prehospital setting, as digital reporting is becoming more common in ambulances. Currently, some NEWS parameters (except for respiratory rate, level of consciousness and body temperature) are already automatically sent to an electronic emergency patient record system in most hospital districts in Finland.[22] Within the next two years, all Finnish EMS systems will have a uniform electronic patient record system. These electronic data could be entered simultaneously during patient evaluation to a machine learning algorithm, which would calculate estimates of short-term (e.g. 24-h, 48-h and 30-day) mortality. These estimates could facilitate EMS staff's recognition of high-risk patients who might otherwise be left at the scene or transported to inappropriate destinations. Future machine learning studies should utilise all available data that are documented in electronic patient record systems, as based on this study, the traditional NEWS parameter and blood glucose combined have a limited potential to predict 30-day mortality.

Strengths and limitations

Our study has limitations attributable to its observational design. First, blood glucose measurements were based on clinical judgement. This introduces selection bias, since these patients were more likely to be higher-risk patients. Nevertheless, only 3.6% of the patients who were otherwise eligible for analysis had an unknown blood glucose level. Second, RF was selected as a machine learning method, although it is unknown which machine learning method is the most suitable for our research question. Third, bootstrapped CIs for AUROC and p-values seem to be conflicting: p-values suggested a statistical significance whereas CIs partly overlapped. Nevertheless, the null hypothesis can be rejected at the $\alpha = 0.05$ level in this kind of scenario.[23] Finally, our RF model's performance should be externally validated in another prospectively collected dataset.

The most noteworthy strengths of the study are related to its design. During the prospective data collection, the ambulance personnel were mandated to measure the standard NEWS parameters in all adult patients encountered, regardless of the type of mission. Additionally, the study population included patients left at the

scene or transported to a general practitioner, which further strengthens model's ability to detect moderate-risk patients among all prehospital patients.

Conclusion

An RF algorithm combining traditional NEWS parameters and blood glucose levels showed a fair performance in predicting 30-day mortality among unselected prehospital patients. Blood glucose improved the RF model's predictive power slightly.

Conflicts of interest

None.

CRediT authorship contribution statement

Joonas Tamminen: Conceptualization, Methodology, Formal analysis, Writing - original draft. **Antti Kallonen:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - original draft, Visualization. **Sanna Hoppu:** Conceptualization, Investigation, Resources, Writing - review & editing, Supervision, Project administration. **Jari Kalliomäki:** Conceptualization, Investigation, Resources, Data curation, Writing - review & editing, Supervision, Project administration.

Acknowledgements

This study was financially supported by the Competitive State Research Financing of the Expert Responsibility area of Tampere University Hospital (Grant 9V006).

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.resplu.2021.100089.

REFERENCES

1. Royal College of Physicians. National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS: Updated report of a working party: Executive summary and recommendations. London: Royal College of Physicians; 2017.
2. Subbe CP. Validation of a modified early warning score in medical admissions. QJM 2001;94:521–6.
3. Singh S, McGlennan A, England A, Simons R. A validation study of the CEMACH recommended modified early obstetric warning system (MEOWS). Anaesthesia 2012;67:12–8.
4. Akre M, Finkelstein M, Erickson M, Liu M, Vanderbilt L, Billman G. Sensitivity of the pediatric early warning score to identify patient deterioration. Pediatrics 2010;125:e763.
5. Schein RMH, Hazday N, Pena M, Ruben BH, Sprung CL. Clinical antecedents to in-hospital cardiopulmonary arrest. Chest 1990;98:1388–92.
6. Hillman KM, Bristow PJ, Chey T, et al. Duration of life-threatening antecedents prior to intensive care admission. Intensive Care Med 2002;28:1629–34.
7. Silcock DJ, Corfield AR, Gowens PA, Rooney KD. Validation of the National Early Warning Score in the prehospital setting. Resuscitation 2015;89:31–5.
8. Patel R, Nugawela MD, Edwards HB, et al. Can early warning scores identify deteriorating patients in pre-hospital settings? A systematic review. Resuscitation 2018;132:101–11.
9. Vihonen H, Lääperi M, Kuisma M, Pirneskoski J, Nurmi J. Glucose as an additional parameter to National Early Warning Score (NEWS) in prehospital setting enhances identification of patients at risk of death: an observational cohort study. Emerg Med J 2020;286–92.
10. Eckart A, Hauser SI, Kutz A, et al. Combination of the National Early Warning Score (NEWS) and inflammatory biomarkers for early risk stratification in emergency department patients: results of a multinational, observational study. BMJ Open 2019;9:1–11.
11. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Crit Care Med. 2016;44:368–74.
12. Ho TK. Random decision forests. Proc 3rd Int Conf Doc Anal Recognit 1995 Aug 14–16; Montreal: IEEE Xplore; 1995. p. 278–282.
13. Statistics Finland. Official statistics of Finland (OSF): Population structure. (Accessed 7 January 2020, at. . https://www.tilastokeskus.fi/tup/suoluk/suoluk_vaesto_en.html).
14. Refaeilzadeh P, Tang L, Liu H. Cross-validation. Encyclopedia of database systems. New York: Springer; 2009. p. 532–8.
15. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347–58.
16. Blomberg SN, Folke F, Ersbøll AK, et al. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Resuscitation 2019;138:322–9.
17. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care 2019;23:1–13.
18. Giannini HM, Ginestra JC, Chivers C, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. Crit Care Med 2019;47:1485–92.
19. Al-Dury N, Ravn-Fischer A, Hollenberg J, et al. Identifying the relative importance of predictors of survival in out of hospital cardiac arrest: a machine learning study. Scand J Trauma Resusc Emerg Med 2020;28:60.
20. Spangler D, Hermansson T, Smekal D, Blomberg H. A validation of machine learning-based risk scores in the prehospital setting. PLoS One 2019;14:1–18.
21. Dungan KM, Braithwaite SS, Preiser JC. Stress hyperglycaemia. Lancet 2009;373:1798–807.
22. Pirneskoski J, Kuisma M, Olkkola KT, Nurmi J. Prehospital National Early Warning Score predicts early mortality. Acta Anaesthesiol Scand 2019;63:676–83.
23. Knezevic A. Overlapping confidence intervals and statistical significance. StatNews 2008. . (Accessed 22 June 2020, at https://cscu.cornell.edu/news/statnews/stnews73.pdf).