Internal and Emergency Medicine
https://doi.org/10.1007/s11739-021-02873-y

EM - ORIGINAL

Machine learning versus traditional methods for the development of risk stratification scores: a case study using original Canadian Syncope Risk Score data

Lars Grant[1,2,3] · Pil Joo[4] · Marie-Joe Nemnom[5] · Venkatesh Thiruganasambandamoorthy[4,5,6]

Received: 17 May 2021 / Accepted: 13 October 2021

Abstract
Artificial Intelligence and machine learning (ML) methods are promising for risk-stratification, but the added benefit over traditional statistical methods remains unclear. We compared predictive models developed using machine learning (ML) methods to the Canadian Syncope Risk Score (CSRS), a risk-tool developed with logistic regression for predicting serious adverse events (SAE) after emergency department (ED) disposition for syncope. We used the prospective multicenter cohort data collected for CSRS development at 11 Canadian EDs over an 8-year period to develop four ML models to predict 30-day SAE (death, arrhythmias, MI, structural heart disease, pulmonary embolism, hemorrhage) after ED disposition. The CSRS derivation and validation cohorts were used for training and testing, respectively, and the 43 variables used included demographics, medical history, vital signs, ECG findings, blood tests and the diagnostic impression of the emergency physician. Performance was assessed using the area under the receiver-operating-characteristics curve (AUC) and calibration curves. Of the 4030 patients in the training set and 3819 patients in the test set overall, 286 (3.6%) patients suffered 30-day SAE. The AUCs for model validation in test data were CSRS 0.902 (0.877–0.926), regularized regression 0.903 (0.877–0.928), gradient boosting 0.914 (0.894–0.934), deep neural network 0.906 (0.883–0.929), simplified gradient boosting 0.904 (0.881–0.927). The AUCs and calibration slopes for the ML models and CSRS were similar. Two ML models used fewer predictors than the CSRS but matched its performance. Overall, the ML models matched the CSRS in performance, with some models using fewer predictors.

Keywords Artificial intelligence · Machine learning · Syncope · Risk stratification · Prediction

Introduction

Background

Machine learning (ML) and artificial intelligence (AI) have recently received significant attention and have been described as the "next major technologic breakthrough to affect health care delivery" [1]. One of the most cited applications is the risk-stratification of patients to predict specific outcomes. Multiple clinical risk scores and decision rules have been developed for emergency department (ED) patients using ML methods [2–11]. On the other hand, risk-stratification had been a focus of emergency medicine research for many years before AI became popular and there exist robust statistical approaches for the development and reporting of risk-tools from prospectively collected data [12, 13]. Many such tools have become a standard part of training and practice in the specialty [14–27].

✉ Venkatesh Thiruganasambandamoorthy
vthirug@ohri.ca

1 Department of Emergency Medicine, McGill University, Montreal, QC, Canada

2 Lady Davis Research Institute, Montreal, QC, Canada

3 Jewish General Hospital, Montreal, QC, Canada

4 The Ottawa Hospital, Ottawa, ON, Canada

5 Clinical Epidemiology Program, Emergency Medicine, Ottawa Hospital Research Institute, 1053 Carling Avenue, Ottawa, ON K1Y 4E9, Canada

6 School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada

Published online: 03 November 2021

Importance

Traditional risk-tool development has been based on a logistic regression applied to prospectively collected research data, typically a few thousand patients. ML methods allow computers to identify and learn complex patterns in large data sets and use such patterns to make predictions in new cases [28] . In principle, ML methods have the potential to outperform logistic regression models by capturing non-linear relationships and interactions among predictors. On the other hand, ML models typically require larger sample sizes for stability, reliability and proper fit because of their increased flexibility [29]. Some prior studies using retrospective data found little benefit with ML methods [30]. It is unclear how beneficial ML models will be in the development of generalizable predictive rules based on prospectively collected research data and as far as we are aware, a direct comparison of an established, high-quality risk stratification tool to a ML based tool derived using the same data set has not yet been attempted.

Goals of this investigation

One risk tool that was previously derived and validated is the Canadian Syncope Risk Score (CSRS; Fig. 1) that predicts the 30-day risk of serious adverse event (SAE; e.g. death, arrhythmia, hemorrhage) after an ED visit for syncope [31, 32]. The objective of this study was to derive and validate modern ML predictive models using the original CSRS data and compare their performance to the traditional logistic regression model, the CSRS.

Category	Points
Clinical Evaluation	
Predisposition to vasovagal symptoms	-1
History of heart disease	1
Any systolic pressure reading <90 or > 180 mm Hg	2
Investigations	
Troponin level elevated > 99% or normal population	2
Abnormal QRS axis (< -30° or > 100°)	1
QRS duration > 130 ms	1
Corrected QT interval > 480 ms	2
Diagnosis in Emergency Department	
Vasovagal syncope	-2
Cardiac syncope	2
Total Score (-3 to 11)	

Fig. 1 Canadian Syncope Risk Score to identify patients with syncope at risk of serious adverse events within 30 days after disposition from the emergency department

Methods

Study setting and population

Our study was a secondary analysis of the original data collected for CSRS development through two prospective cohort studies detailed below [31, 32].

Derivation and validation of the CSRS

The derivation phase of the CSRS enrolled 4322 adult patients who presented to one of six Canadian EDs (a list of participating EDs is shown in Appendix 1) within 24 h of syncope from September 2010 to February 2014. After excluding those who had a serious condition identified during the index ED evaluation and those lost to 30-day follow-up, the CSRS was derived with data from 4030 patients. The tool was derived to predict any one of the listed SAE (Appendix 2) within 30-days of ED disposition. A total of 147 (3.6%) patients suffered 30-day SAE.

An initial list of 43 candidate predictors (Table 1) were considered, of which 23 variables were selected for multivariable logistic regression by excluding those with sparse distribution, large proportion of missingness, unacceptable interobserver agreement values and non-significance on bivariate analysis. The final model included nine predictors and optimism corrected area under the receiver-operating-characteristic curve (AUC) was 0.87 (95% CI 0.84–0.89). The regression coefficients were translated into scores.

The validation phase of the CSRS [32] enrolled adult syncope patients at nine Canadian EDs between March 2014 to April 2018. After exclusions, 3819 ED patients were included in the final analysis and 139 (3.6%) patients experienced 30-day SAE. The AUC of the CSRS in the validation cohort was 0.91 (95% CI 0.88–0.93). The calibration slope was 1.0 indicating accurate risk estimation for patients in the validation cohort.

Data used to develop the ML models

We used the same 4030 patients in the derivation phase to train and 3819 patients in the validation set to test the ML models. All 43 predictor variables that were considered for CSRS derivation (Table 1) were included. The CSRS derivation excluded many predictors prior to analysis based on selection criteria as described above, but these predictors were included in our ML analysis because ML methods are designed to handle many predictors. As in the CSRS

Table 1 List of 43 predictors considered for inclusion in the CSRS

24 categorical predictors
Sex
6 Syncope characteristics: Witnessed, Palpitations prior to the syncope, syncope while sitting or lying or exertion, predisposition to vasovagal symptoms (warm-crowded place, prolonged standing, fear, emotion or pain), presence of prodrome (dizziness, light-headedness, vision changes, nausea or vomiting), or presence of orthostatic symptoms prior or after the syncope
2 Medical history predictors: heart disease (history of any one of the following: coronary or valvular heart disease, cardiomyopathy, congestive heart failure or non-sinus rhythm—ECG evidence during the index visit or documented history of ventricular or atrial arrhythmias, or device implantation); or vascular disease (history of transient ischemic attack, cerebrovascular accident or peripheral vascular disease)
2 Family history predictors: congenital heart disease or sudden death
2 Final ED diagnosis predictors at disposition: vasovagal or cardiac with other diagnosis as the reference variable and includes cause unknown
Troponin > 99th percentile of the normal population
10 ECG predictors: 5 blocks (right bundle branch block, left bundle branch block, left anterior fascicular block, left posterior fascicular block, bifascicular block [right bundle branch block + left fascicular block either anterior or posterior, or bundle branch block + first degree atrioventricular block); axis deviations—left or right; ventricular hypertrophy—right or left; or presence of old ischemia
19 Continuous Predictors
Age
11 ED vitals: Triage, highest and lowest systolic and diastolic blood pressures; triage, highest and lowest pulse rates, triage respiratory rate and oxygen saturation
4 laboratory values: hemoglobin, hematocrit, urea and creatinine levels
3 ECG predictors: QRS axis, QRS duration, and corrected QT interval

This is the same list of predictors that were used in our ML models

development, troponin levels were dichotomized at the 99th percentile threshold for the normal population.

The outcome to be predicted by the ML models was the same 30-day composite outcome after ED disposition used for the CSRS development which included any of the following serious adverse events: death, arrhythmia, myocardial infarction, serious structural heart disease, aortic dissection, pulmonary embolism, severe pulmonary hypertension, severe hemorrhage, subarachnoid hemorrhage and any other serious condition causing syncope and procedural interventions for the treatment of syncope (such as pacemaker insertion). More details about the composite outcome are available in Appendix 2.

Data analysis

We used descriptive analysis with mean, standard deviation (SD), and range for continuous variables and proportions for categorical variables [31, 32]. Correlations between the 43 predictor variables were assessed using Spearman correlation coefficients.

Where a patient's age was missing, it was imputed as the median age of the training data set. The categorical variables in Table 1, such as troponin and ECG findings, were imputed based on the assumption that if data for one of these variables were missing, the condition described by that variable was absent. This is the same imputation strategy that was used in the CSRS derivation and validation phases. The other continuous variables in Table 1 were imputed by linear regression based on the age and sex of each patient. The regression models used for this imputation were fit on the training data only and then applied to both the training and the test data.

We developed four competing ML models that use the predictors to provide an estimated probability of the outcome: a regularized regression model, a gradient boosting model, a simplified gradient boosting model using only a small number of predictors and a deep neural network model.

Details of ML modeling

The regularized regression model was developed in two stages. First, to handle potential correlations among the 43 predictors, a least absolute shrinkage and selection operator (LASSO) regularized logistic regression was used for variable selection. Second, a ridge regularized logistic regression model was trained using only the predictors selected by the LASSO regression. Class weights inversely proportional to the frequencies of the outcome were used because of the imbalanced nature of the data set. The gradient boosting model used regression trees with deviance loss. The gradient boosting and deep neural network models used all 43 predictors. The deep neural network model contained three fully connected hidden layers with rectified linear unit (RELU) activation, binary cross-entropy loss and an Adam [33] optimizer. A dropout layer was included for improved regularization. All the ML models were trained using either

Platt's or an isotonic calibration method. A grid search was performed using threefold cross validation to choose optimal hyperparameters.

The simplified gradient boosting model was produced by performing backward predictor selection, sequentially reducing the number of predictors in the gradient boosting model by dropping the least important predictors until the test-set AUC began to drop significantly. The performance characteristics of the resulting simplified model were reported along with the other models described above.

Assessment of ML model performance

The relative importance of the predictors for each of the models was assessed using beta coefficients for the regularized logistic regression. Permutation feature importance [34] was used for the gradient boosting and deep neural network models. To deal with multicollinearity, clusters of predictors with correlation coefficients > 0.45 were defined and the least important predictors were sequentially removed from each of these clusters to identify a single representative predictor in each cluster before performing a permutation feature importance analysis with the rest of the predictors.

The ML models were compared to the CSRS using four approaches: receiver-operator characteristic curve (ROC) analysis, calibration curves, Brier scores, and decision curve analysis (DCA). We used 5000 bootstrap samples from test data to calculate the Brier score, and the AUC with 95% confidence intervals and t-tests to assess the difference between the mean AUCs for the different models. We generated calibration curves and used a linear fit to assess the slope and intercept for each of the models. The Brier score is sensitive to both the discriminatory power of a model and the quality of its calibration. To compare the potential clinical implications of using the various models, a decision curve analysis (DCA) was carried out [35, 36].

Results

There were 4030 patients in the training set and 3819 patients in the test set. Overall, 286 (3.6%) of these patients suffered the study outcome (Table 2).

We identified five clusters of predictors with Spearman correlation coefficient > 0.45 as shown in Appendix 3. There were no other moderate or strong correlations among the 43 predictors. The proportions of patients with data missing in the training and test datasets are detailed in Appendix 4. Only three variables had a missingness of > 25% in the training set (family history of congenital heart disease, family history of sudden death and a troponin level > 99 percentile of the normal population). For each of these categorical variables, a missing value is far more likely to represent an absence of the condition than vice-versa. All missing data were imputed as detailed in the methods section.

ROC curves for the CSRS and the ML models are shown in Fig. 2. Table 3 shows the performance characteristics of the models. Figure 2 suggests that the gradient

Table 2 Characteristics, emergency department management and outcomes of patients in the training set (the CSRS derivation set [31]) and the test set (the CSRS validation set [32])

Variable	Training set (CSRS derivation set) No. (%) of patients $n = 4030$	Test set (CSRS validation set) No. (%) of patients $n = 3819$
Age in year, mean ± SD [range]	53.6 ± 23.0 [16–102]	53.9 ± 22.8 [16–101]
Female	2238 (55.5)	2088 (54.7)
Arrival by ambulance	2590 (64.3)	2396 (62.7)
Medical history		
Hypertension	1292 (32.1)	1113 (29.1)
Diabetes mellitus	402 (10.0)	424 (11.1)
Coronary artery disease	476 (11.8)	397 (10.4)
Atrial fibrillation or flutter	291 (7.2)	243 (6.4)
Valvular heart disease	137 (3.4)	115 (3.0)
Congestive heart failure	152 (3.8)	90 (2.4)
Management in emergency department		
Electrocardiography	3834 (95.1)	3705 (97.0)
Blood tests	3446 (85.5)	3091 (80.9)
Hospitalized	381 (9.5)	335 (8.8)
Outcome		
Serious adverse event during index visit hospitalization	86 (2.1)	85 (2.2)
Serious adverse event after the index visit	61 (1.5)	54 (1.4)

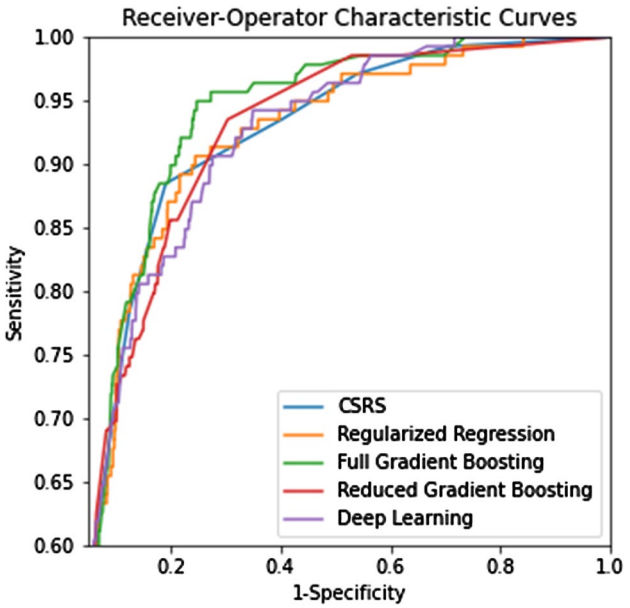

Fig. 2 Comparison of receiver-operator characteristic curves for machine learning models and the CSRS

boosting model exhibited improved discrimination compared to the CSRS in a range of risk-thresholds where the sensitivity is 0.90–0.97, but that all models perform similarly when the sensitivity is required to be higher than 0.97. There was no statistically significant difference in the overall AUC among the ML models and the CSRS.

Calibration curves for the models are shown in Appendix 5. Slopes of the calibration curves, as extracted using linear regression, are shown in Table 3. All models had calibration slopes close to 1 with intercepts near 0 and high adjusted R^2. The Brier scores of all models are shown in Table 3. All models were reasonably well calibrated. Because the outcome occurs in only 3.6% of the cases in the test set, calibration assessments are sensitive to the choice of how probabilities are binned when creating the calibration curves and for this reason we do not report t-test results for comparisons between the Brier scores of the different models.

Decision curves are shown in Appendix 6 and suggest that there would be a potential benefit to using the gradient boosting model if the true ideal risk-threshold for intervention (admission or consultation in the ED) is in the range 5–20%. Otherwise, the curves of the ML models and the CSRS were similar.

The nine most important predictors for each model are shown in Table 4. Many of the most important predictors were similar between the models. The regularized regression model and the simplified gradient boosting model use only seven and eight predictors, respectively. rather than the nine predictors used by the CSRS but had similar performance characteristics in the test set.

Discussion

Based on the same data used to derive and validate the CSRS, we produced four models using modern ML methods to predict 30-day SAE after ED disposition. To our knowledge, this study is the first to directly compare a validated traditional risk stratification tool and ML models derived using the same prospective dataset.

The ML models matched the performance of the CSRS in all global performance measures and we observed a trend toward improved performance under certain conditions. Decision Curve Analysis suggests an increased net clinical benefit with use of the ML models as compared to the CSRS if the ideal risk-threshold for ED specialist consultation or admission lies between 5 and 20%. Additionally, two of the ML models (the regularized regression model and the simplified gradient boosted decision tree model) matched the performance of the CSRS despite using fewer predictors.

In summary, our work shows that ML methods can match the performance of the CSRS and two of our ML models do so using fewer predictors than the CSRS. These findings suggest that ML methods are a viable alternative to traditional approaches in the development of clinical decision rules

Table 3 Performance characteristics of the models in the test set

	Area under the curve (95% CI)	Brier score (95% CI)	Slope [intercept] (adjusted R^2) of calibration curve	*t* statistic*	*p* value*
Canadian syncope risk score model	0.902 (0.877–0.926)	0.029 (0.025–0.034)	1.36 [−0.03] (0.980)	Reference	Reference
Regularized regression	0.903 (0.877–0.928)	0.028 (0.024–0.032)	1.08 [−0.02] (0.973)	0.075	0.94
Gradient boosting	0.914 (0.894–0.934)	0.029 (0.025–0.033)	0.93 [−0.00] (0.902)	1.496	0.13
Deep neural network	0.906 (0.883–0.929)	0.029 (0.024–0.033)	1.28 [−0.02] (0.959)	0.651	0.51
Simplified gradient boosting	0.904 (0.881–0.927)	0.030 (0.026–0.034)	0.99 [0.01] (0.955)	0.193	0.85

*Two-sided *t* tests for area under the curve (AUC) comparison to the Canadian Syncope Risk Score model

Table 4 Top 9 most important predictors in each model

CSRS	LR	Simplified GB	GB	DL
Troponin > 99% of the normal population	ED diagnosis of cardiac syncope	ED diagnosis of cardiac syncope	ED diagnosis of vasovagal syncope	ED diagnosis of vasovagal syncope
ED diagnosis of cardiac syncope	Heart disease history	ED diagnosis of vasovagal syncope	Troponin > 99% of the normal population	Heart disease history
ED diagnosis of vasovagal syncope	ED diagnosis of vasovagal syncope	QTc > 480 ms	ED diagnosis of cardiac syncope	Troponin > 99% of the normal population
QTc > 480 ms	QTc	Troponin > 99% of the normal population	Heart disease history	ED diagnosis of cardiac syncope
Systolic blood pressure < 90 or > 180 mmHg	Troponin > 99% of the normal population	QRS duration	Age	Age
QRS duration > 130 ms	QRS duration	Hematocrit	QRS duration	Predisposition to vasovagal symptoms
Predisposition to vasovagal symptoms	Bifascicular block	Age	Hemoglobin	Presence of orthostatic symptoms
Heart disease history		Lowest ED SBP	QTc	Right bundle branch block
Abnormal QRS axis			QRS axis	QTc

The CSRS and LR model predictors were sorted according to the absolute value of their beta coefficients. The GB and DL models were analysed using permutation feature importance to rank the predictors

GB gradient boosting, *DL* deep neural network, *LR* regularized regression

and support the increased use of ML in clinical decision rules research. ML models and traditional models can be developed in parallel and comparisons between the models can inform the choice of predictors and lead to an improved predictive tool, whether it be a ML tool or a traditional risk stratification score.

One key role of ML in the development of risk stratification tools may be in variable selection. Our simplified gradient boosted decision trees and regularized regression models illustrate how ML can process a large number of predictor variables impartially and produce a model with only a few, most relevant inputs. This can sidestep subjective variable selection procedures that can involve many clinicians and dissenting opinions thereby saving time and resources in the development process. For example, we note that five of the predictors selected by ML were common to the simplified gradient boosting, regularized regression and CSRS models: ED diagnosis of cardiac or vasovagal syncope, abnormal troponin levels, QTc and QRS duration. The remaining predictors chosen varied by model and included a history of heart disease, presence of bifascicular block, age, hematocrit and lowest ED systolic blood pressure. Each of these variables has good face value from a clinical perspective and can be used in bedside clinical reasoning: if all these predictors are favorable in a given case, the patient in question is very likely to be low risk. Furthermore, the CSRS involves three predictors that involve the subjective judgment of the ED clinician: ED diagnosis of cardiac or vasovagal syncope and predisposition to vasovagal syncope. The regularized regression syncope model eliminated one of these subjective predictors, predisposition to vasovagal syncope. It also used two fewer predictors but achieved the same performance as the CSRS.

Given that ML methods generally require large training data sets to achieve optimal performance [29], we expect that predictive ML modeling will demonstrate an increased benefit when using larger and more complex data. Cohorts consisting of more than a few thousand cases can be difficult to obtain in a prospective research context but have become readily available in an observational context using electronic health records (EHR). Additionally, EHRs frequently contain many relevant predictor variables, and, in this context, ML methods have the capacity to detect non-linear effects and complex interactions that would be missed by standard analysis. The derivation of risk-stratification tools using EHR data will likely be an important application of ML, but difficulties with non-stationarity and generalizability are important challenges to this approach [37–39].

The predictive performance of ML models should not be considered in isolation, but along side other factors such as interpretability and portability. The deep neural network model may be the least attractive of the models presented in our work given that it is more difficult to interpret. Explainable and interpretable AI methods may be needed to mitigate the perceived black-box nature of some ML algorithms [40] whereas traditional models such as the CSRS are inherently easy to interpret and apply at the bedside. Regarding portability, many modern EHR systems will allow the integration of ML algorithms such as those developed in our study directly into the electronic workflow of care providers thereby increasing their usability and our regularized regression and simplified gradient

boosting models use a smaller number of predictors than the CSRS and could easily be deployed as simple online risk calculators like those already available for the CSRS at MDCalc [41] and elsewhere.

A common weakness in risk scoring systems is the incorporation of subjective physician impression as a predictor variable. The CSRS, for example, includes ED physician diagnostic impression of cardiac or vasovagal syncope as a predictor. Several other commonly used risk scoring rules also use subjective predictors [16, 26, 42]. In the future, methods of ML, possibly in combination with natural language processing may eliminate the need for such subjective variables by objectively identifying specific features in the history of present illness that best predict the outcome.

The strength of our study is the direct comparison of the ML and traditional statistical models developed using the same high-quality dataset.

Limitations include the fact that our statistical model comparison is based on the AUC which is a global metric. Partial AUC analysis would allow for a statistical comparison of the models in specific threshold ranges of the ROC curve that may be of clinical interest. Additionally, the size of the training data set may not have allowed the ML models to reach their optimal performance potential as large data sets are ideal for the training of ML models [29]. Finally, machine learning modeling might perform better or worse in other data sets of similar size using other outcomes because ML performance depends not only on the size of the data sets, but also other factors such as data quality and the presence of non-linear effects or complex interactions.

Conclusion

Our ML models matched the performance of the CSRS and some models did so while using fewer predictors than the CSRS.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11739-021-02873-y.

Author contributions LG, PJ and VT conceived the idea. All authors contributed to the study design and developed the study protocol. VT applied for funding. LG, PJ and M-JN provided guidance on data management. LG drafted the manuscript. All authors reviewed the manuscript and contributed substantially to its revision. LG takes responsibility for the paper as a whole.

Funding Previously collected data for two prospective studies were used in the study. The two prospective studies were funded by The Physicians' Services Incorporated Foundation (09q4017), Canadian Institutes of Health Research (MOP-114927), Heart and Stroke Foundation Canada (G-15-0009006), and the Cardiac Arrhythmia Network of Canada (SRG-15-P10-001) as part of the Networks of Centres of Excellence (NCE).

Availability of data and material My manuscript has no associated data.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no competing interests. At the time of the research, Dr. Thiruganasambandamoorthy held a salary award 'National New Investigator Award' through the Heart and Stroke Foundation of Canada. Currently, he holds the Physicians' Services Incorporated (PSI) Foundation's 'PSI-50 Mid-Career Clinical Research Award'. The funders take no responsibility for the design, conduct, results or interpretations presented here.

Ethics approval Ethical approval was obtained for the two prospective cohort studies from Ottawa Health Science Network Research Ethics Board; derivation protocol #2009661-01H and the validation protocol #20160137-01H.

Human and animal rights The two prospective cohort studies included human participants but not animals.

Consent to participate The ethics committees approved the study with the requirement of only verbal consent.

References

1. Grant K, McParland A, Mehta S, Ackery AD (2020) Artificial intelligence in emergency medicine: surmountable barriers with revolutionary potential. Ann Emerg Med 75(6):721–726. https://doi.org/10.1016/j.annemergmed.2019.12.024
2. Graham B, Bond R, Quinn M, Mulvenna M (2018) Using data mining to predict hospital admissions from the emergency department. IEEE Access 6:10458–10469. https://doi.org/10.1109/ACCESS.2018.2808843
3. Lee EK, Yuan F, Hirsh DA, Mallory MD, Simon HK (2012) A clinical decision tool for predicting patient care characteristics: patients returning within 72 hours in the emergency department. AMIA Annu Symp Proc 2012:495–504
4. Harrison RF, Kennedy RL (2005) Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. Ann Emerg Med 46(5):431–439. https://doi.org/10.1016/j.annemergmed.2004.09.012
5. Eken C, Bilge U, Kartal M, Eray O (2009) Artificial neural network, genetic algorithm, and logistic regression applications for predicting renal colic in emergency settings. Int J Emerg Med 2(2):99–105. https://doi.org/10.1007/s12245-009-0103-1
6. Sun Y, Heng BH, Seow YT, Seow E (2009) Forecasting daily attendances at an emergency department to aid resource planning. BMC Emerg Med 9:1. https://doi.org/10.1186/1471-227X-9-1
7. Taylor RA, Pare JR, Venkatesh AK et al (2016) Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. Acad Emerg Med 23(3):269–278. https://doi.org/10.1111/acem.12876
8. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K (2019) Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care 23(1):64. https://doi.org/10.1186/s13054-019-2351-7
9. Liu N, Zhang Z, Wah Ho AF, Hock Ong ME (2018) Artificial intelligence in emergency medicine. J Emerg Crit Care Med.

http://jeccm.amegroups.com/article/view/4700. Accessed 25 Sept 2020
10. Adedinsewo D, Carter RE, Attia Z et al (2020) Artificial intelligence-enabled ECG algorithm to identify patients with left ventricular systolic dysfunction presenting to the emergency department with dyspnea. Circ Arrhythm Electrophysiol 13(8):e008437. https://doi.org/10.1161/CIRCEP.120.008437
11. Jang D-H, Kim J, Jo YH et al (2020) Developing neural network models for early detection of cardiac arrest in emergency department. Am J Emerg Med 38(1):43–49. https://doi.org/10.1016/j.ajem.2019.04.006
12. Collins GS, Reitsma JB, Altman DG, Moons Karel GM (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). Circulation 131(2):211–219. https://doi.org/10.1161/CIRCULATIONAHA.114.014508
13. Stiell IG, Wells GA (1999) Methodologic standards for the development of clinical decision rules in emergency medicine. Ann Emerg Med 33(4):437–447. https://doi.org/10.1016/s0196-0644(99)70309-4
14. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR (1992) A study to develop clinical decision rules for the use of radiography in acute ankle injuries. Ann Emerg Med 21(4):384–390. https://doi.org/10.1016/s0196-0644(05)82656-3
15. Stiell IG, Greenberg GH, McKnight RD et al (1993) Decision rules for the use of radiography in acute ankle injuries. Refinement and prospective validation. JAMA 269(9):1127–1132. https://doi.org/10.1001/jama.269.9.1127
16. Six AJ, Backus BE, Kelder JC (2008) Chest pain in the emergency room: value of the HEART score. Neth Heart J 16(6):191–196
17. Backus BE, Six AJ, Kelder JC et al (2010) Chest pain in the emergency room: a multicenter validation of the HEART Score. Crit Pathw Cardiol 9(3):164–169. https://doi.org/10.1097/HPC.0b013e3181ec36d8
18. Hoffman JR, Wolfson AB, Todd K, Mower WR (1998) Selective cervical spine radiography in blunt trauma: methodology of the National Emergency X-Radiography Utilization Study (NEXUS). Ann Emerg Med 32(4):461–469. https://doi.org/10.1016/s0196-0644(98)70176-3
19. Hoffman JR, Mower WR, Wolfson AB, Todd KH, Zucker MI (2000) Validity of a set of clinical criteria to rule out injury to the cervical spine in patients with blunt trauma. National Emergency X-Radiography Utilization Study Group. N Engl J Med 343(2):94–99. https://doi.org/10.1056/NEJM200007133430203
20. Kline JA, Mitchell AM, Kabrhel C, Richman PB, Courtney DM (2004) Clinical criteria to prevent unnecessary diagnostic testing in emergency department patients with suspected pulmonary embolism. J Thromb Haemost 2(8):1247–1255. https://doi.org/10.1111/j.1538-7836.2004.00790.x
21. Kline JA, Courtney DM, Kabrhel C et al (2008) Prospective multicenter evaluation of the pulmonary embolism rule-out criteria. J Thromb Haemost 6(5):772–780. https://doi.org/10.1111/j.1538-7836.2008.02944.x
22. Kuppermann N, Holmes JF, Dayan PS et al (2009) Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. Lancet 374(9696):1160–1170. https://doi.org/10.1016/S0140-6736(09)61558-0
23. Schonfeld D, Bressan S, Da Dalt L, Henien MN, Winnett JA, Nigrovic LE (2014) Pediatric Emergency Care Applied Research Network head injury clinical prediction rules are reliable in practice. Arch Dis Child 99(5):427–431. https://doi.org/10.1136/archdischild-2013-305004
24. Stiell IG, Wells GA, Vandemheen K et al (2001) The Canadian CT Head Rule for patients with minor head injury. Lancet 357(9266):1391–1396. https://doi.org/10.1016/s0140-6736(00)04561-x
25. Stiell IG, Clement CM, Rowe BH et al (2005) Comparison of the Canadian CT Head Rule and the New Orleans Criteria in patients with minor head injury. JAMA 294(12):1511–1518. https://doi.org/10.1001/jama.294.12.1511
26. Wells PS, Anderson DR, Rodger M et al (2001) Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and d-dimer. Ann Intern Med 135(2):98–107. https://doi.org/10.7326/0003-4819-135-2-200107170-00010
27. Wolf SJ, McCubbin TR, Feldhaus KM, Faragher JP, Adcock DM (2004) Prospective validation of Wells Criteria in the evaluation of patients with suspected pulmonary embolism. Ann Emerg Med 44(5):503–510. https://doi.org/10.1016/j.annemergmed.2004.04.002
28. Rajkomar A, Dean J, Kohane I (2019) Machine learning in medicine. N Engl J Med 380(14):1347–1358. https://doi.org/10.1056/NEJMra1814259
29. van der Ploeg T, Austin PC, Steyerberg EW (2014) Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 14(1):137. https://doi.org/10.1186/1471-2288-14-137
30. Pua Y-H, Kang H, Thumboo J et al (2020) Machine learning methods are comparable to logistic regression techniques in predicting severe walking limitation following total knee arthroplasty. Knee Surg Sports Traumatol Arthrosc 28(10):3207–3216. https://doi.org/10.1007/s00167-019-05822-7
31. Thiruganasambandamoorthy V, Kwong K, Wells GA et al (2016) Development of the Canadian Syncope Risk Score to predict serious adverse events after emergency department assessment of syncope. CMAJ 188(12):E289–E298. https://doi.org/10.1503/cmaj.151469
32. Thiruganasambandamoorthy V, Sivilotti MLA, Sage NL et al (2020) Multicenter emergency department validation of the Canadian syncope risk score. JAMA Intern Med 180(5):737–744. https://doi.org/10.1001/jamainternmed.2020.0288
33. Kingma DP, Ba J (2017) Adam: a method for stochastic optimization. http://arxiv.org/abs/1412.6980 [cs]. Accessed 13 Jan 2021
34. Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res 20(177):1–81
35. Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 26(6):565–574. https://doi.org/10.1177/0272989X06295361
36. Vickers AJ, van Calster B, Steyerberg EW (2019) A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. https://doi.org/10.1186/s41512-019-0064-7
37. Jung K, Shah NH (2015) Implications of non-stationarity on predictive modeling using EHRs. J Biomed Inform 58:168–174. https://doi.org/10.1016/j.jbi.2015.10.006
38. Nestor B, McDermott MBA, Boag W et al (2019) Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. http://arxiv.org/abs/1908.00690 [cs, stat]. Accessed 26 Dec 2020
39. Callahan A, Shah NH, Chen JH (2020) Research and reporting considerations for observational studies using electronic health record data. Ann Intern Med 172(11_Supplement):S79–S84. https://doi.org/10.7326/M19-0873

40. Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052
41. Canadian Syncope Risk Score (2020) MDCalc. https://www.mdcalc.com/canadian-syncope-risk-score. Accessed 27 Dec 2020
42. Wells PS, Anderson DR, Rodger M et al (2003) Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis. N Engl J Med 349(13):1227–1235. https://doi.org/10.1056/NEJMoa023153