Saffari *et al.*
BMC Medical Research Methodology (2022) 22:286
https://doi.org/10.1186/s12874-022-01770-y

BMC Medical Research Methodology

RESEARCH

Open Access

AutoScore-Ordinal: an interpretable machine learning framework for generating scoring models for ordinal outcomes

Seyed Ehsan Saffari[1,2†], Yilin Ning[1†], Feng Xie[1,2], Bibhas Chakraborty[1,2,3,4], Victor Volovici[5,6], Roger Vaughan[1,2], Marcus Eng Hock Ong[2,7] and Nan Liu[1,2,8,9*]

Abstract

Background: Risk prediction models are useful tools in clinical decision-making which help with risk stratification and resource allocations and may lead to a better health care for patients. AutoScore is a machine learning–based automatic clinical score generator for binary outcomes. This study aims to expand the AutoScore framework to provide a tool for interpretable risk prediction for ordinal outcomes.

Methods: The AutoScore-Ordinal framework is generated using the same 6 modules of the original AutoScore algorithm including variable ranking, variable transformation, score derivation (from proportional odds models), model selection, score fine-tuning, and model evaluation. To illustrate the AutoScore-Ordinal performance, the method was conducted on electronic health records data from the emergency department at Singapore General Hospital over 2008 to 2017. The model was trained on 70% of the data, validated on 10% and tested on the remaining 20%.

Results: This study included 445,989 inpatient cases, where the distribution of the ordinal outcome was 80.7% alive without 30-day readmission, 12.5% alive with 30-day readmission, and 6.8% died inpatient or by day 30 post discharge. Two point-based risk prediction models were developed using two sets of 8 predictor variables identified by the flexible variable selection procedure. The two models indicated reasonably good performance measured by mean area under the receiver operating characteristic curve (0.758 and 0.793) and generalized c-index (0.737 and 0.760), which were comparable to alternative models.

Conclusion: AutoScore-Ordinal provides an automated and easy-to-use framework for development and validation of risk prediction models for ordinal outcomes, which can systematically identify potential predictors from high-dimensional data.

Keywords: Interpretable machine learning, Medical decision making, Clinical score, Ordinal outcome, Electronic health records

Introduction

Risk prediction models are mathematical equations which help clinicians estimate the probability of a healthcare outcome, given patient data. Such models include integer-point scores which can be used to predict that a disease is present (diagnostic models) or a specific outcome will occur (prognostic models), depending on the clinical question. A combination of multiple predictors

†Seyed Ehsan Saffari and Yilin Ning contributed equally to this work.

*Correspondence: liu.nan@duke-nus.edu.sg

[1] Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore
Full list of author information is available at the end of the article

(different weights for different predictors) is included into a multivariable model to calculate a risk score [1–3]. Some risk prediction models have been used in routine clinical settings, including the Framingham Risk Score [4], Ottawa Ankle Rules [5], Nottingham Prognostic Index [6], Gail model [7], Euro-SCORE [8], the modified Early Warning Score (MEWS) [9, 10] and Simplified Acute Physiology Score [11].

The use of health information technology, particularly electronic health records (EHR), has increased in the past decade, which provides opportunities for big data research. EHR data includes detailed patient information and clinical outcome variables which can be a unique data source for risk model development [12, 13]. Availability of a large number of variables in EHR data could be a mathematical challenge when using traditional regression analysis to build up a risk model. Machine learning (ML), as an alternative approach, applies mathematical algorithms to handle such big data resulting in novel risk prediction models. Traditional variable selection approaches (such as backward elimination, forward selection, stepwise selection using pre-specified stopping rules) may result in different subsets of variables in the context of EHR data, and clinical knowledge might not be always available in some clinical domains. Powerful feature selection techniques are available for supervised learning, which is a very critical aspect in risk model development when working with EHR data [13, 14].

AutoScore [15] is an easy-to-use, machine learning–based automatic clinical score generator, which develops interpretable clinical scoring models. In an empirical experiment using EHR data, AutoScore generated scoring models that achieved comparable predictive performance as several conventional methods for risk model development but by using fewer variables [15]. The advantage of the AutoScore framework is the combination of efficient variable selection using ML techniques and the accessibility and interpretability of simple regression models. It can be easily used in different clinical settings and its applicability has been shown with a large number of variables (EHR data, for example) [15]. Some recent studies have used this framework to develop a risk prediction model in various clinical domains [16–20].

Most risk prediction models in the literature were developed using multivariable logistic regression models or ML techniques to predict a binary outcome. Aside from the AutoScore framework, ML applications include the use of Naive Bayes (NB), XGBoost, k-nearest neighbor (K-NN), multilayer perceptron, support vector machine (SVM) and CatBoost for predicting the risk of cardiovascular disease [21], random forest (RF), XGBoost, logistic regression, SVM and K-NN for the risk of incident diabetic retinopathy among patients with type 2 diabetes mellitus [22], a stroke risk prediction model using NB, decision tree and RF models [23], a XGBoost based cerebral infarction risk prediction model [24], and a developed risk model for 90-day mortality of patients undergoing gastric cancer resection with curative intent using cross validated elastic regularized logistic regression method, boosting linear regression, RF and an ensemble model [25].

Many clinical ordinal outcome variables exist and they are often dichotomized (favorable and unfavorable) or reduced to unordered categories for simplicity, e.g., in a cross-sectional study of emergency department (ED) triage [26] and a retrospective cohort study of ovarian cancer patients [27]. Nevertheless, it should not be ignored that such re-categorization results in loss of clinically and statistically relevant information, which may also involve difficulties in borderline patients (cases that can easily be categorized into either of the two levels of the outcome). One should note that analyzing ordinal variables has more statistical power in comparison to the corresponding re-categorized binary variables. This has been illustrated in both simulations and empirical studies in clinical trials [28–32]. Literature also recommends the use of the ordinal scale outcomes rather than dichotomization, as smaller treatment effect sizes are detectable via ordinal analysis [29, 33–35].

In the literature, ordinal outcome variables are discussed in several clinical domains, where the objective was either an association exploration or predictions. A large international study (including 26 hospitals from six countries) conducted ordinal logistic regression to study a composite ordinal outcome variable (defined as 1 = alive, no long length of stay [LOS], no readmission; 2 = alive, long LOS, no readmission; 3 = alive, no long LOS, readmission; 4 = alive, long LOS, readmission; 5 = death), and the correlation among different levels of the composite ordinal outcome at hospital level was reported [36]. ML methods using multiple biomarkers were performed to develop an ovarian cancer–specific predictive framework in a retrospective cohort study of 435 patients on a secondary ordinal outcome of residual tumor size (defined as: no residual tumor, <1 cm residual tumor, ≥ 1 cm residual tumor), and the predictive accuracy and AUC were discussed [27]. Statistical and ML methods have been used for ordinal outcomes in the literature, e.g., the proportional odds model (POM) in middle ear dysfunction diagnosis of infants [37] and in a coronary artery disease study [38], ordinal RF in the aforementioned ovarian cancer study [27], multilayer perceptron with ordinal loss in a study across 9 mental health and suicide-related sub-Reddits [39], and 3D convulotional neural network model with ordinal binary decomposition in Parkinson's disease

patients [40]. However, there is a lack of interpretability (where one may not easily understand the output of such complex and how it works, which is not recommended in healthcare domain [41]) and accessibility using these ML approaches, whereas the transparent POM is not as easily used as an interpretable risk scoring system in the clinic for real-time decision making.

There is a lack of literature in model development using ordinal analysis that can be easily applied to clinical studies dealing with complex data (EHR, for example). The primary objective of this study was to expand the original AutoScore framework to provide a tool for easy development and validation of risk prediction models for ordinal outcomes. Hence the main contribution of the current study is not only the inclusion of the ordinal blocks, but also some modifications on the original AutoScore framework which leads to new methodological work and revised model performance measurements appropriate for ordinal outcomes. For illustration purpose, a risk prediction model was developed and validated using EHR data from the emergency department (as a real world data), where the ordinal outcome included three categories (alive without readmission to the hospital within 30 days post discharge, alive with readmission within 30 days post discharge and dead inpatient or within 30 days post discharge).

Methods

AutoScore-Ordinal framework

In this section we describe the 6 modules constituting the proposed AutoScore-Ordinal framework. In Module 1 (see Fig. 1) the data is first split into a training set to train prediction models, a validation set to select hyperparameters (e.g., number of variables, cut-off values for categorizing continuous variables), and a test set to evaluate the final model(s) selected. The three datasets typically contain 70%, 10% and 20% of the full dataset, respectively. Variables are ranked based on their importance to a RF [42] for multiclass classification (i.e., ignoring the ordering of categories), trained on the training set with a default number of 100 trees.

To simplify the interpretation and account for possible non-linear relationship between the predictor variables and the outcome, all continuous variables are categorized in Module 2 (see Fig. 1). To automate this process, AutoScore-Ordinal categorizes each continuous variable using the 5-th, 20-th, 80-th and 95-th percentiles (based on the training set) as cut-off values, but some cut-offs may be removed to avoid sparsity issues when the distribution of a variable is highly skewed. These (somewhat arbitrary) cut-off values provide reasonable initial configuration for subsequent score development, and can be fine-tuned by users in Module 5 (see detail below).

In Module 3 (see Fig. 1), weights associated with variables are developed using the cumulative link model

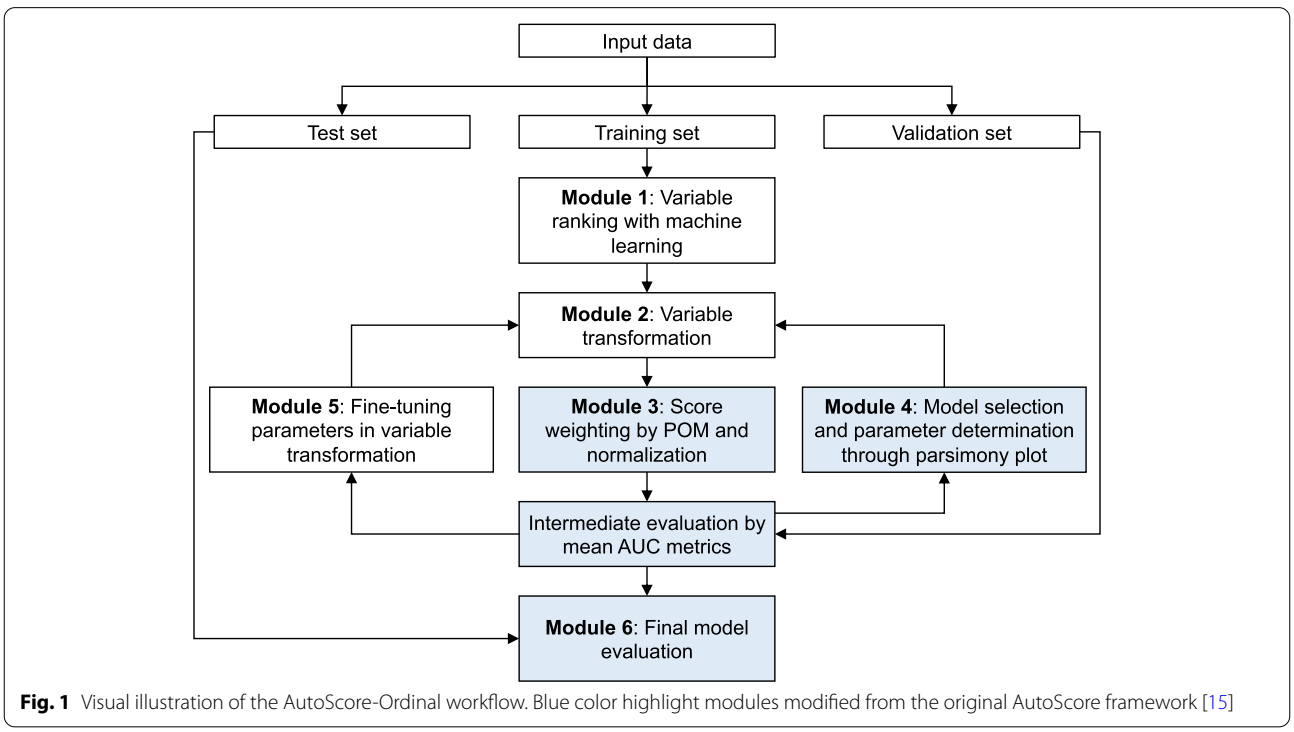


Fig. 1 Visual illustration of the AutoScore-Ordinal workflow. Blue color highlight modules modified from the original AutoScore framework [15]

[43] with the logit link, also known as the proportional odds model (POM) [43, 44], which is one of the most widely used regression models in studies of ordinal outcomes and has been integrated with deep learning approaches to handle complex (e.g., image) data [45]. Let scalar Y denote the ordinal outcome with J categories (denoted by integers 1, ..., J) and column vector $\boldsymbol{x}$ denote the variables (with continuous variables readily categorized in Module 2). The POM assumes a linear model for the logit of the cumulative probabilities associated with the j-th ordinal category, i.e., $p_j = P(Y \leq j)$, $j = 1, ..., J-1$:

$$\log\left(\frac{p_j}{1-p_j}\right) = \theta_j - \boldsymbol{x}^T\boldsymbol{\beta}.$$

The scalar terms θ_j are category-specific intercept terms, where $\theta_1 < \theta_2 < ... < \theta_{J-1}$ to ensure $p_j < p_k$ for any $j < k$. $\boldsymbol{\beta}$ is the vector of regression coefficients corresponding to the predictors. The negative sign before $\boldsymbol{\beta}$ follows from the notation used by McCullagh [43, 44], such that a positive value of $\boldsymbol{\beta}$ indicates a positive association between $\boldsymbol{x}$ and Y, i.e., an increase in $\boldsymbol{x}$ leads to an increased probability of observing a higher category in Y. Hence an increase in $\boldsymbol{x}^T\boldsymbol{\beta}$ is always associated with increased probabilities of observing higher outcome categories, allowing us to construct prediction scores based on $\boldsymbol{x}^T\boldsymbol{\beta}$. Another general approach for handling ordinal outcomes is ordinal binary decomposition, but it models an ordinal outcomes as several binary labels in separate models [46], making it challenging to derive a common score for the risk of being in each ordinal category.

A simple scaling and rounding of trained $\boldsymbol{\beta}$ values may generate a scoring model spanning negative and positive values with confusing interpretation, e.g., the arbitrary zero score may be misinterpreted as no risk. Hence, the POM is refitted after redefining reference categories in each variable such that all elements in $\boldsymbol{\beta}$ are positive, and $\boldsymbol{\beta}$ is normalised with respect to the minimum value of $\boldsymbol{\beta}$. With all continuous variables readily categorised in Module 2, these normalised coefficients can be interpreted as scores associated with a category of a variable, referred to as partial scores. The partial scores (which are 0 for reference categories and 1 or larger otherwise) are rounded to positive integers to simplify the calculation of final prediction scores, which is the summation of all partial scores corresponding to the values of variables for an individual. To facilitate interpretation, all partial scores are often rescaled (and then rounded) such that the maximum total score attenable is a meaningful value (e.g., 100).

To evaluate the performance of the final model, the prediction of outcome Y with J categories is divided into $J-1$ binary classifications of $Y \leq j$ vs $Y > j$, and the mean area under the receiver operating characteristic curve (AUC) across these binary classifications (referred to as mAUC hereafter) is used to evaluate the overall performance for predicting Y, which is equivalent to the average dichotomized c-index for evaluating ordinal predictions [47, 48]. In Module 4, a scoring model is grown by adding one variable at each time (based on the variable ranking from Module 1) until all candidate variables are included, and the improvement in mAUC (evaluated on the validation set) with increasing number of variables is inspected using the parsimony plot. The final list of variables is often selected when the benefit of adding a variable is small, where such small benefit could be assessed via visual inspection (by looking at parsimony plot) and clinical knowledge (and drop/include variables manually). Next, the cut-off values for continuous variables selected in Module 4 may be fine-tuned for favourable interpretation in Module 5, e.g., by using 10-year age groups instead of the arbitrarily defined quantile-based intervals. The final model is evaluated on the test set in Module 6 using the mAUC and Harrell's generalised c-index [47, 49, 50], which is based on the proportion of concordant pairs (i.e., when predictions and observed outcomes generate the same ranking for the pair of observations, including tied ranks) among all possible pairs of observations. For both mAUC and generalised c-index, a value of 0.5 indicates a random performance and a value of 1 indicates a perfect predictive performance. The mAUC and generalised c-index from the test set are reported with the bias-corrected 95% bootstrap confidence interval (CI) [51].

Data preparation

To demonstrate and validate our proposed AutoScore-Ordinal framework, we applied it in a clinical study in compliance with the checklist for assessment of medical AI [52]. We used AutoScore-Ordinal to predict readmission and death (composite outcome) after inpatient discharge, using data collected from patients who visited the emergency department (ED) of Singapore General Hospital in years 2008 to 2017 and were subsequently admitted to the hospital [53, 54]. The full cohort included data on 449,593 ED presentation cases. Information on patient demographics, ED administration, inpatient admission, clinical tests and vital signs in ED, medical history and comorbidities was extracted from the hospital electronic health record system [16]. We excluded patients aged below 18, resulting in a final sample of 445,989 inpatient cases.

We constructed a composite ordinal outcome with three categories: alive without readmission to the

hospital within 30 days post discharge, alive with readmission within 30 days post discharge, died inpatient or within 30 days post discharge. Among the 445,989 cases, 359,961 (80.7%) were in the first outcome category (i.e., alive without 30-day readmission), 55,552 (12.5%) were in the second category (i.e., alive with 30-day readmission), and 30,476 (6.8%) were in the third category (i.e., died inpatient or by day 30 post discharge).

We randomly split the dataset (stratified by outcome categories) into a training set of 70% ($n=312,193$) cases to train models, a validation set of 10% ($n=44,599$) cases to perform necessary model fine-tuning for AutoScore-Ordinal, and a test set of 20% ($n=89,197$) cases to evaluate the performance of the final prediction models. For each case, we extracted the length of stay (LOS) of the previous inpatient admission (missing values were treated as 0 days). Missing values for vital signs or clinical tests were imputed using the median value in the validation set.

We compared the prediction model built using AutoScore-Ordinal with the RF (with 100 trees) and POM with LASSO or stepwise variable selection techniques. For each model, we computed the 95% CI for mAUC and generalized c-index from bootstrap samples of the test set (the number of bootstrap samples was selected as 100 for the demo purposes and can be modified in the AutoScore algorithm). Generalized c-index was computed based on the total score for AutoScore-generated models, the linear predictor excluding intercept terms for POM and the predicted outcome categories for RF.

Implementation

All analyses were implemented in R version 4.0.5 [55]. Our proposed AutoScore-Ordinal is implemented as an R package, available from https://github.com/nliulab/AutoScore-Ordinal. POM was implemented using the *clm* function from package *ordinal* [56]. The *stepAIC* function from package *MASS* [57] was used to perform stepwise variable selection for POM, and the *ordinalNet* function from package *ordinalNet* [58] was used to implement the LASSO approach. The RF was implemented using the *randomForest* function from package *randomForest* [59]. The bias-corrected bootstrap CI was implemented using the *bca* function from package *coxed* [60]. The generalized c-index was implemented using the *rcorrcens* function from package *Hmisc* [61].

Results

The characteristics of the full cohort are summarized in Table 1. Cases in the 3 outcome categories showed statistical difference in all variables, therefore it is non-trivial to develop a sparse prediction model based on POM.

Variable selection

The parsimony plot (see Fig. 2) suggests a reasonable model of the first 8 variables: ED LOS, creatinine, ED boarding time, number of visits in the previous year, age, systolic blood pressure (SBP), bicarbonate and pulse, which reached a mAUC that is only 7.9% lower than that the scoring model using all 41 variables. We refer to this model as Model 1. When using the parsimony plot to select variables, researchers are not restricted to consecutively select variables in the descending order of importance. For example, we built an alternative model (i.e., Model 2) with 8 variables, where we excluded the 3rd variable (i.e., ED boarding time) from Model 1 that had little impact on mAUC, and added the 14th variable (i.e., history of metastatic cancer in the past 5 years, which can be easily collected by asking the patient or the accompanying person/family/relatives) that incremented the mAUC by approximately 4% when it entered the prediction model.

Fine-tuning

All variables selected in the two models were continuous, and we fine-tuned their cut-off values in the categorization step to improve interpretability. The scoring tables after fine-tuning were shown in Table 2 for both models, and the performance of the resulting prediction models (evaluated on the test set) were reported in Table 3. Model 1 had an mAUC of 0.758 (95% CI: 0.754–0.762), and by excluding ED boarding time and adding metastatic cancer, the mAUC of Model 2 improved to 0.793 (95% CI: 0.789–0.796).

Interpreting prediction scores

The AutoScore-generated score (from Models 1 and 2) can be mapped to the likelihood of falling into different outcome categories based on the observed proportions in the training set. For example, we illustrate the use of Model 2 for risk prediction for a hypothetical new patient in Fig. 3. With values of the 8 variables measured for this new patient, clinicians can simply check relevant rows in the scoring table, summate the partial scores to a total score for this patient, and read the corresponding predicted probabilities for the three outcome categories in the lookup table. Such predicted probabilities can also be calculated from POM using a calculator or be returned from RF using designated software commands, but the checklist-style scoring table of AutoScore-generated models and the accompanying lookup tables of predicted probabilities are much easier to use in clinical practice.

We evaluate the calibration performance of Models 1 and 2, visually presented in Fig. 4. Specifically, we grouped subjects based on score intervals defined in the lookup table in Fig. 3, and plotted the observed risk of

Table 1 Characteristics of cases in the full cohort. Outcome categories 1, 2, and 3 refer to cases that were alive without readmission to the hospital within 30 days post discharge, alive with readmission within 30 days post discharge and dead inpatient or within 30 days post discharge, respectively

	Overall ($n = 445,989$)	**Outcome category 1 (alive, no readmission; $n = 359,961$)**	**Outcome category 2 (alive, with readmission; $n = 55,552$)**	**Outcome category 3 (death; $n = 30,476$)**
Patient demographics				
Age (years; mean (SD))	61.66 (18.24)	60.16 (18.55)	66.38 (15.86)	70.84 (13.83)
Male (%)	222,644 (49.9)	177,267 (49.2)	28,753 (51.8)	16,624 (54.5)
Race (%)				
Chinese	27,471 (6.2)	24,615 (6.8)	1958 (3.5)	898 (2.9)
Indian	316,474 (71.0)	250,930 (69.7)	41,022 (73.8)	24,522 (80.5)
Malay	47,508 (10.7)	39,606 (11.0)	5973 (10.8)	1929 (6.3)
Others	54,536 (12.2)	44,810 (12.4)	6599 (11.9)	3127 (10.3)
Comorbidity (%)				
Myocardial infarction	26,594 (6.0)	15,653 (4.3)	6242 (11.2)	4699 (15.4)
Congestive heart failure	49,575 (11.1)	32,360 (9.0)	11,809 (21.3)	5406 (17.7)
Peripheral vascular disease	25,878 (5.8)	16,701 (4.6)	6258 (11.3)	2919 (9.6)
Stroke	57,730 (12.9)	41,674 (11.6)	10,463 (18.8)	5593 (18.4)
Dementia	12,385 (2.8)	8129 (2.3)	2625 (4.7)	1631 (5.4)
Pulmonary	42,770 (9.6)	30,385 (8.4)	8868 (16.0)	3517 (11.5)
Rheumatic	6180 (1.4)	4645 (1.3)	1147 (2.1)	388 (1.3)
Peptic ulcer disease	17,193 (3.9)	11,834 (3.3)	3478 (6.3)	1881 (6.2)
Mild liver disease	20,483 (4.6)	14,318 (4.0)	4216 (7.6)	1949 (6.4)
Severe liver disease	7119 (1.6)	3863 (1.1)	1906 (3.4)	1350 (4.4)
Diabetes (without complications)	55,699 (12.5)	42,529 (11.8)	8756 (15.8)	4414 (14.5)
Diabetes with complications	104,682 (23.5)	76,553 (21.3)	19,987 (36.0)	8142 (26.7)
Paralysis	24,903 (5.6)	17,683 (4.9)	4692 (8.4)	2528 (8.3)
Renal	91,213 (20.5)	62,033 (17.2)	20,290 (36.5)	8890 (29.2)
Cancer (non-metastatic)	39,571 (8.9)	27,627 (7.7)	6778 (12.2)	5166 (17.0)
Metastatic cancer	35,225 (7.9)	18,469 (5.1)	5683 (10.2)	11,073 (36.3)
ED Admission				
ED LOS (hours; mean (SD))	2.86 (1.70)	2.84 (1.72)	2.58 (1.62)	2.12 (1.42)
ED Triage code (%)				
P1	83,221 (18.7)	59,513 (16.5)	11,696 (21.1)	12,012 (39.4)
P2	250,382 (56.1)	199,708 (55.5)	33,906 (61.0)	16,768 (55.0)
P3 and P4	112,386 (25.2)	100,740 (28.0)	9950 (17.9)	1696 (5.6)
ED Boarding Time (hours; mean (SD))	4.78 (3.81)	4.80 (3.79)	4.79 (3.94)	4.48 (3.89)
Consultation Waiting Time (hours; mean (SD))	0.77 (0.80)	0.80 (0.82)	0.71 (0.72)	0.53 (0.60)
Inpatient admission				
Day of Week (%)				
Friday	62,453 (14.0)	50,314 (14.0)	7801 (14.0)	4338 (14.2)
Monday	74,192 (16.6)	60,142 (16.7)	9091 (16.4)	4959 (16.3)
Weekend	115,418 (25.9)	92,387 (25.7)	14,604 (26.3)	8427 (27.7)
Midweek	193,926 (43.5)	157,118 (43.6)	24,056 (43.3)	12,752 (41.8)
Admission Type (%)				
A1	16,814 (3.8)	14,795 (4.1)	1195 (2.2)	824 (2.7)
B1	37,345 (8.4)	32,938 (9.2)	2658 (4.8)	1749 (5.7)
B2	212,261 (47.6)	174,752 (48.5)	23,238 (41.8)	14,271 (46.8)
C	179,569 (40.3)	137,476 (38.2)	28,461 (51.2)	13,632 (44.7)
Previous LOS (days; mean (SD))	3.57 (8.55)	3.04 (7.90)	5.34 (10.00)	6.50 (11.55)

Table 1 (continued)

	Overall ($n = 445,989$)	Outcome category 1 (alive, no readmission; $n = 359,961$)	Outcome category 2 (alive, with readmission; $n = 55,552$)	Outcome category 3 (death; $n = 30,476$)
Healthcare utilisation in the previous year				
No. inpatient visits (mean (SD))	0.93 (2.21)	0.62 (1.42)	2.66 (4.46)	1.44 (2.17)
No. surgery (mean (SD))	0.20 (0.74)	0.15 (0.63)	0.42 (1.10)	0.37 (0.99)
No. ICU stays (mean (SD))	0.02 (0.25)	0.02 (0.22)	0.05 (0.35)	0.05 (0.36)
No. HD stays (mean (SD))	0.09 (0.47)	0.07 (0.40)	0.17 (0.68)	0.17 (0.69)
Vital sign and clinical tests				
Ventilation (%)	89 (0.0)	47 (0.0)	7 (0.0)	35 (0.1)
Resuscitation (%)	9083 (2.0)	6211 (1.7)	1045 (1.9)	1827 (6.0)
Pulse, beat/minute (mean (SD))[a]	82.85 (17.23)	81.99 (16.73)	83.35 (17.00)	92.05 (20.51)
Respiration, breath/minute (mean (SD))[a]	17.84 (1.77)	17.76 (1.60)	17.98 (1.81)	18.59 (2.98)
SpO_2, % (mean (SD))[a]	97.98 (3.25)	98.05 (3.05)	97.93 (3.12)	97.34 (5.22)
DBP, mmHg (mean (SD))[a]	71.33 (13.55)	71.67 (13.36)	71.14 (13.70)	67.62 (14.93)
SBP, mmHg (mean (SD))[a]	133.68 (25.53)	134.05 (25.19)	135.98 (26.22)	125.13 (26.63)
Bicarbonate, mmol/L (mean (SD))[a]	22.78 (3.78)	22.91 (3.55)	22.55 (3.97)	21.67 (5.39)
Creatinine, µmol/L (mean (SD))[a]	154.70 (208.52)	142.72 (196.69)	213.21 (259.28)	185.61 (215.82)
Potassium, mmol/L (mean (SD))[a]	4.16 (0.72)	4.14 (0.69)	4.23 (0.76)	4.36 (0.91)
Sodium, mmol/L (mean (SD))[a]	135.02 (5.17)	135.30 (4.84)	134.56 (5.39)	132.76 (7.30)

Outcome categories were compared using Kruskal-Wallis and Chi-square test for continuous and categorical variables, respectively. All tests had p-value< 0.001

DBP diastolic blood pressure, *ED* emergency department, *HD* high dependency ward, *ICU* intensive care unit, *LOS* length of stay, *SBP* systolic blood pressure, *SD* standard deviation, SpO_2 blood oxygen saturation

[a] Excluding 9365 missing entries for pulse, 10,772 missing entries for respiration, 10,704 missing entries for SpO_2, 5348 missing entries for SBP and DBP, 56857 missing entries for bicarbonate, 56,742 missing entries for creatinine, 58,747 missing entries for potassium, and 56,678 missing entries for sodium

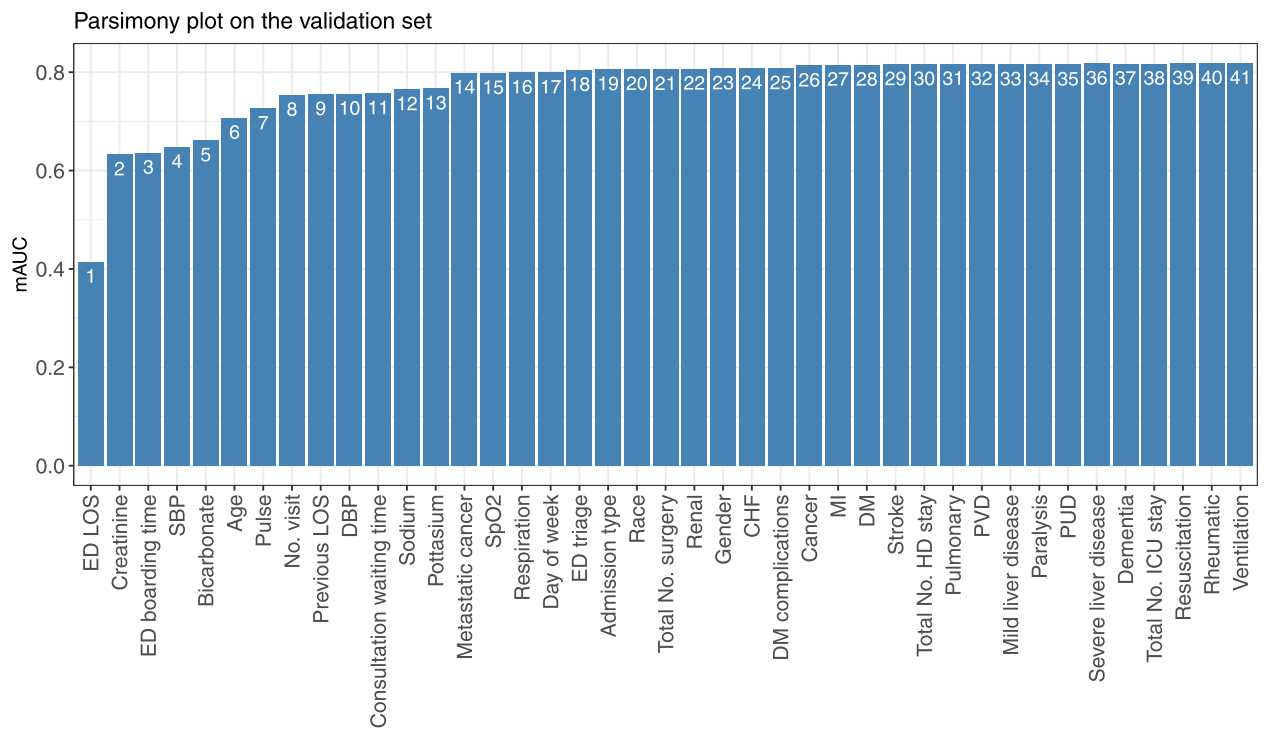


Fig. 2 Parsimony plot by the mean area under the curve (mAUC) on the validation set

Table 2 Scoring table for AutoScore-generated models

Variable	Interval	Partial score for Model 1	Partial score for Model 2
ED LOS	< 40 min	11	7
	[40 min, 80 min)	8	6
	[80 min, 4 h)	4	2
	[4 h, 6 h)	1	1
	> = 6 h	0	0
Creatinine, μmol/L	< 45	6	4
	[45, 60)	0	0
	[60, 135)	0	1
	[135, 595)	6	7
	> = 595	4	7
ED boarding time	< 80 min	0	–
	[80 min, 2.5 h)	2	–
	> = 2.5 h	1	–
Systolic blood pressure, mmHg	< 100	12	9
	[100, 110)	7	5
	[110, 150)	3	2
	[150, 180)	1	0
	> = 180	0	0
Bicarbonate, mmol/L	< 17	8	7
	[17, 20)	3	3
	[20, 28)	0	0
	> = 28	5	4
Age, years	< 25	0	0
	[25, 45)	6	5
	[45, 75)	18	13
	[75, 85)	22	17
	> = 85	25	21
Pulse, beat/minute	< 70	0	0
	[70, 95)	3	2
	[95, 115)	8	6
	> = 115	14	11
Number of inpatient visits in the previous year	< 1	0	0
	[1, 4)	12	9
	> = 4	23	20
Metastatic cancer	No	–	0
	Yes	–	19

"[A, B)" indicates an interval inclusive of the lower limit and exclusive of the upper limit. "--" indicates variables not included in a model

h hours, *min* minutes, *ED* Emergency department, *LOS* length of hospital stay

being in each outcome category in the test set against the predicted risk (based on the lookup tables). Both Models 1 and 2 generated predicted risk similar to observed levels, indicated by dots close to the diagonal line. An increase in the scores (visually indicated by lighter color in Fig. 4) generally reflects an increased likelihood of being in a higher category in the outcome, whereas Model 2 has improved ability compared to Model 1 in differentiating different outcome categories given different predicted scores (indicated by a wider spread of dots along the diagonal line).

Comparison with other approaches

AutoScore-generated prediction models had comparable mAUC as the POM that used the same variables (see Table 3, where POM1 and POM2 correspond to Models 1 and 2 respectively). The RF using the same variables as Model 1 (see RF1 in Table 3) had a higher mAUC than Model 1, but when compared with Model 2 the advantage of the corresponding RF (see RF2 in Table 3) in terms of mAUC is less pronounced. AutoScore-generated models had slightly higher generalized c-index than the corresponding POMs, and both were higher than the corresponding RFs. In particular, the generalized c-index of RFs were much lower than the corresponding AutoScore-generated models or POMs, due to the use of predicted labels instead of numeric scores when evaluating the performance of RF.

When using traditional model building methods to build sparse POM, stepwise algorithm using AIC failed to work when starting from the null model (i.e., without any variable), and ended up selecting 35 variables when starting from the full model (i.e., including all 41 variables). Although this POM with 35 models had a high mAUC and generalized c-index (see POM (stepwise) in Table 3), it is difficult to use in practical settings. The LASSO approach selected 10 variables (i.e., ED LOS, gender, ED triage code, total number of ICU stays in the past year, admission type, SpO_2, SBP, bicarbonate, sodium and diabetes with complications) that had much lower performance than other models (see POM (LASSO) in Table 3).

Discussion

A scoring system was developed using the AutoScore framework for ordinal outcomes in this study. The algorithm was applied on a case study to discuss the risk prediction model and its application on EHR data from the emergency department where the ordinal outcome includes three categories (alive without readmission to the hospital within 30 days post discharge, alive with readmission within 30 days post discharge and dead inpatient or within 30 days post discharge). The model was developed using 70% of the data ($n = 312{,}193$); validated on subset of 10% of the data ($n = 44{,}599$) to perform necessary model fine-tuning; and tested on a set of 20% ($n = 89{,}197$). The performance of the AutoScore-Ordinal model was checked against the alternative models including POM and RF using 100 bootstrap samples via mAUC and generalized c-index. The AutoScore-Ordinal identified two feasible scoring models with 8 variables,

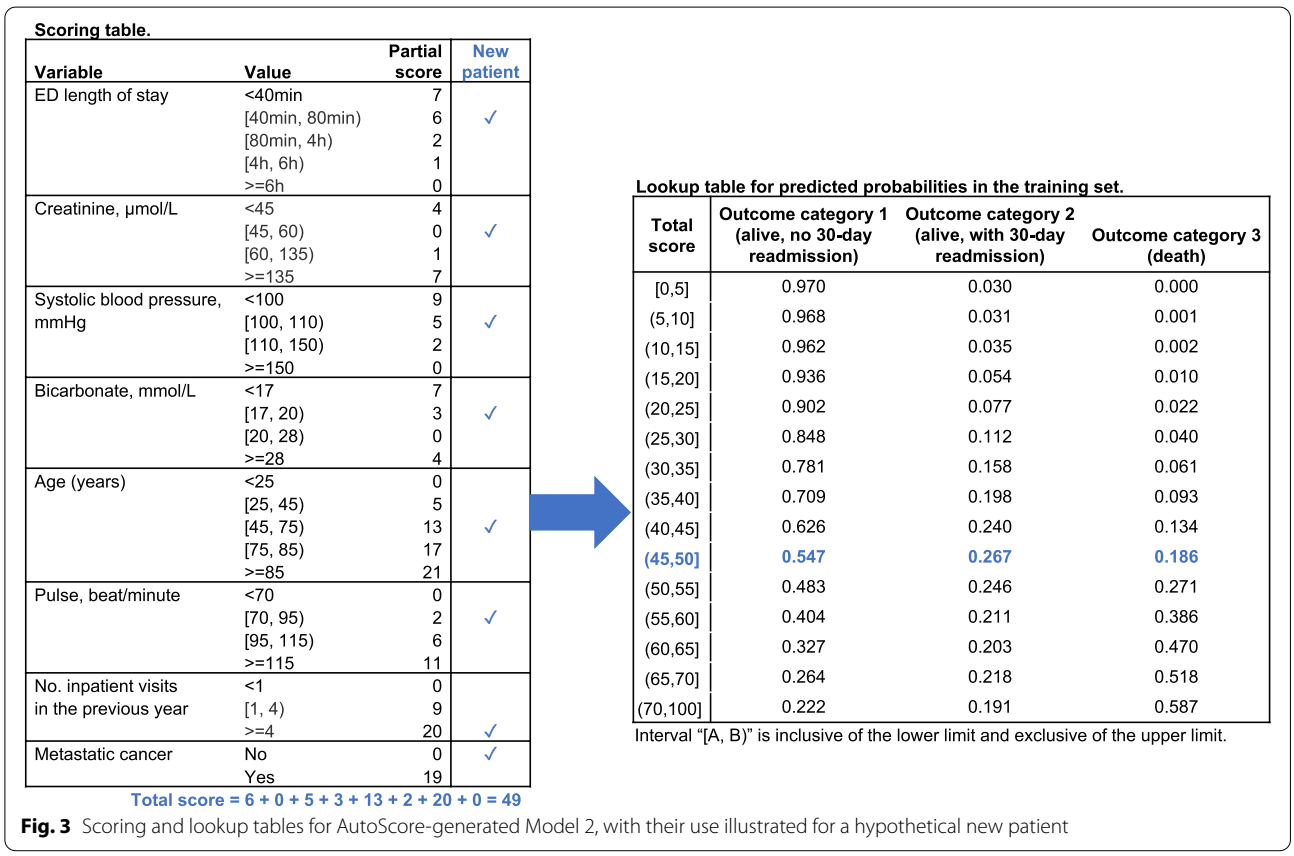
Table 3 Evaluation of prediction models on the test set, after fine-tuning cut-off values for continuous variables. The 95% CIs were generated from 100 bootstrap samples of the test set

	Number of variables	mAUC (95% CI)	Generalized c-index (95% CI)
AutoScore-Ordinal Model 1[a]	8	0.758 (0.754, 0.762)	0.737 (0.734, 0.741)
POM1[a]	8	0.750 (0.747, 0.754)	0.726 (0.722, 0.729)
RF1[a]	8	0.767 (0.764, 0.771)	0.547 (0.544, 0.549)
AutoScore-Ordinal Model 2[b]	8	0.793 (0.789, 0.796)	0.760 (0.757, 0.763)
POM2[b]	8	0.790 (0.786, 0.793)	0.754(0.750, 0.756)
RF2[b]	8	0.798 (0.794, 0.801)	0.564 (0.561, 0.566)
POM (stepwise)	35	0.815 (0.812–0.819)	0.775 (0.772–0.778)
POM (LASSO)	10	0.704 (0.700–0.708)	0.669 (0.665–0.673)

POM proportional odds model, *RF* random forest, *mAUC* mean area under the curve

[a] These models used the same 8 variables: emergency department (ED) length of stay (LOS), creatinine, ED boarding time, number of visits in the previous year, age, systolic blood pressure (SBP), bicarbonate and pulse

[b] These models used the same 8 variables: ED LOS, creatinine, number of visits in the previous year, age, SBP, bicarbonate, pulse and metastatic cancer



Scoring table.

Variable	Value	Partial score	New patient
ED length of stay	<40min	7	
	[40min, 80min)	6	✓
	[80min, 4h)	2	
	[4h, 6h)	1	
	>=6h	0	
Creatinine, μmol/L	<45	4	
	[45, 60)	0	✓
	[60, 135)	1	
	>=135	7	
Systolic blood pressure, mmHg	<100	9	
	[100, 110)	5	✓
	[110, 150)	2	
	>=150	0	
Bicarbonate, mmol/L	<17	7	
	[17, 20)	3	✓
	[20, 28)	0	
	>=28	4	
Age (years)	<25	0	
	[25, 45)	5	
	[45, 75)	13	✓
	[75, 85)	17	
	>=85	21	
Pulse, beat/minute	<70	0	
	[70, 95)	2	✓
	[95, 115)	6	
	>=115	11	
No. inpatient visits in the previous year	<1	0	
	[1, 4)	9	
	>=4	20	✓
Metastatic cancer	No	0	✓
	Yes	19	

Total score = 6 + 0 + 5 + 3 + 13 + 2 + 20 + 0 = 49

Lookup table for predicted probabilities in the training set.

Total score	Outcome category 1 (alive, no 30-day readmission)	Outcome category 2 (alive, with 30-day readmission)	Outcome category 3 (death)
[0,5]	0.970	0.030	0.000
(5,10]	0.968	0.031	0.001
(10,15]	0.962	0.035	0.002
(15,20]	0.936	0.054	0.010
(20,25]	0.902	0.077	0.022
(25,30]	0.848	0.112	0.040
(30,35]	0.781	0.158	0.061
(35,40]	0.709	0.198	0.093
(40,45]	0.626	0.240	0.134
(45,50]	0.547	0.267	0.186
(50,55]	0.483	0.246	0.271
(55,60]	0.404	0.211	0.386
(60,65]	0.327	0.203	0.470
(65,70]	0.264	0.218	0.518
(70,100]	0.222	0.191	0.587

Interval "[A, B)" is inclusive of the lower limit and exclusive of the upper limit.

Fig. 3 Scoring and lookup tables for AutoScore-generated Model 2, with their use illustrated for a hypothetical new patient

and both had slightly better performance than the POM and RF that use the same variables. The novelty of the AutoScore-Ordinal model is its easy-to-use and machine learning-based automatic clinical score generator features, which develops interpretable clinical scoring models and can be useful tools for clinical decision-making at different stages of clinical pathway.

Prediction models in clinical settings are useful tools to inform clinical decision-making at different stages of clinical practice [62, 63]. To design, conduct and build prediction models, fundamental concepts including developing, validating and updating risk prediction models are discussed in the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual

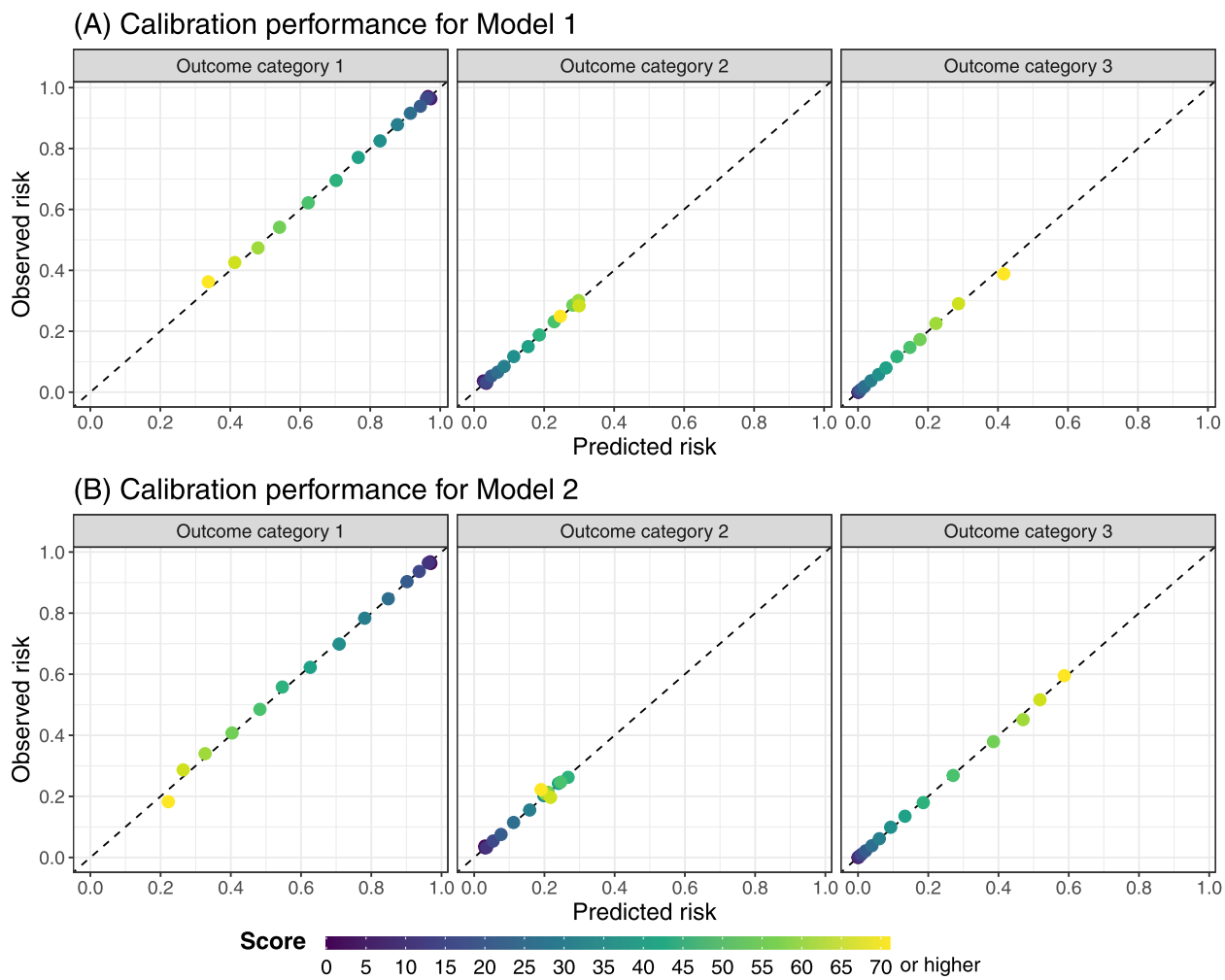


Fig. 4 Calibration performance for (**A**) Model 1 and (**B**) Model 2

Prognosis Or Diagnosis) Statement [64]. New risk models should always be validated to quantify the predictive ability of the model (for example, calibration and discrimination), which could be addressed via internal (bootstrapping, cross-validation, etc.) or external (independent cohort, for example) validation [64].

Most of the developed models in literature lacks of interpretability and accessibility while using machine learning techniques [26, 27, 39]. In contrast, the AutoScore-Ordinal via a point-based risk prediction model can be easily implemented in different clinical settings and fills a gap in interpretability, when dealing with ordinal outcomes. The advantages of the original AutoScore framework [15] applies to the AutoScore-Ordinal framework. AutoScore-Ordinal builds on the POM, which is suitable for analyzing ordinal outcomes and widely used in clinical and epidemiological research. Compared to conventional use of POM, AutoScore-Ordinal makes use of machine learning methods to build sparse prediction models with good prediction performance, whereas traditional approaches such as stepwise variable selection and LASSO may not work well. AutoScore-Ordinal creates a check-list style scoring model that is easily implemented in clinical settings. In clinical research, sometimes quantitative data are categorized as ordinal variables due to different reasons such as skewness or multi-modal distribution. Under such scenarios, dichotomization may not be ideal and could result in loss of clinically and statistically relevant information. One may take advantage of the AutoScore-Ordinal framework to deal with such ordinal outcome variables.

AutoScore-Ordinal provides an efficient, straightforward and flexible variable selection procedure based on the parsimony plot, which visually presents the improvement in model performance with a growing number of variables in the model. Intuitively, researchers can

select the top few variables that correspond to a satisfying model performance and inclusion of an additional variable results in a small (e.g., <1%) improvement, which resulted in Model 1 in our example. In addition, AutoScore-Ordinal allows researchers to manually add or remove variables from the final variables based on their contribution to model performance (e.g., as illustrated in Model 2) or practical implications. While the current AutoScore-Ordinal implementation uses the POM (or more generally the cumulative link model with the logit link) that is widely used in clinical applications, it can be used with other link functions (e.g., probit, complementary log-log) with minor modifications for possible improvements in model fit. Researchers may want to draw multiple parsimony plots to select a link function that best suits the data when determining variables to include in the final model.

In our data example we trained RF with 100 trees when ranking variables in Module 1 of AutoScore-Ordinal and when using it as a prediction model. Researchers may want to increase the number of trees to improve performance in general applications, e.g., 500 trees is a common choice [65]. Due to the large sample size of our case study, we run out of memory when training an RF with 500 trees, and an RF with 200 trees generated comparable results when ranking variables and predicting ordinal outcomes.

As indicated by the name, POM assumes proportional odds, i.e., the effect of each variable on the outcome is the same across outcome categories. In univariable POM analyses of the training set (without categorizing continuous variables), the proportional odds assumption was rejected for all variables (with significance level of 5%). Future study should investigate how to relax this assumption when necessary without considerably complicating the interpretation of the resulting scoring model. Despite this, the two prediction models built using AutoScore-Ordinal worked reasonably well. For performance evaluation, we considered two metrics (i.e., mAUC and generalized c-index) that have straightforward interpretation and similar definition with metrics for binary and survival predictions [47, 48, 50]. Future work may consider other performance metrics, e.g., volume under the receiver operating characteristic surface (more generally the hypervolume under the manifold) [66] and the ordinal c-index [47] for ordinal prediction, or the M-index [67] and polytomous discrimination index [68, 69] for multi-class outcomes without explicitly accounting for ordering of categories.

Our data example aims to illustrate the use of our proposed AutoScore-Ordinal framework. The prediction performance can be improved, e.g., although Model 2 had better performance than Model 1, it will most likely fail to predict any new case into category 2, as this category is dominated by the other two categories (see lookup table in Fig. 3). The AutoScore-Ordinal should be applied in other clinical domains with different sample sizes and various number of variables to establish external validity. Further investigation is required to improve performance before applying the AutoScore-Ordinal-derived scoring models in clinical settings, e.g., inclusion of additional relevant variables, alternative imputation of missing values and cross-validation feature within the package. Another future research direction, as seen in the literature [70–73], is to integrate the AutoScore-Ordinal package as a mobile application where it could be easily accessible to the clinicians. Nonetheless, AutoScore-Ordinal provides a powerful, flexible and easy-to-use framework for developing interpretable scoring models for ordinal clinical outcomes.

Conclusion

AutoScore-Ordinal as a risk prediction model was developed for ordinal outcome variable. For illustration purpose, the framework was implemented and validated using EHR data from the emergency department, where the ordinal outcome included three categories (alive without readmission to the hospital within 30 days post discharge, alive with readmission within 30 days post discharge and dead inpatient or within 30 days post discharge). An efficient and flexible variable selection procedure was explained and the model indicated a comparable goodness-of-fit in compared to the alternative models. The point-based risk prediction model generated by the AutoScore-Ordinal is easy to implement and interpret in different clinical settings.

Acknowledgements

None.

Authors' contributions

NL: study conception and design, supervision and mentorship. ES, YN and FX: model development, first draft write-up. ES and YN: data analysis. ES, YN, FX, BC, VV, RV, MO, and NL: substantial contributions to results interpretation, algorithm improvement, and critical revision of the manuscript. All authors have reviewed the results, read and approved the final version of the manuscript.

Funding

This study was supported by Duke-NUS Medical School, Singapore. YN is supported by the Khoo Postdoctoral Fellowship Award (project no. Duke-NUS-KPFA/2021/0051) from the Estate of Tan Sri Khoo Teck Puat. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The datasets of this study are not publicly available but available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
This study was approved by Singapore Health Services' Centralized Institutional Review Board (CIRB 2021/2122), and a waiver of consent was granted for EHR data collection. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
None.

Author details
[1]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore. [2]Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore. [3]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA. [4]Department of Statistics and Data Science, National University of Singapore, Singapore, Singapore. [5]Department of Neurosurgery, Erasmus MC University Medical Center, Rotterdam, The Netherlands. [6]Department of Public Health, Erasmus MC, Rotterdam, The Netherlands. [7]Department of Emergency Medicine, Singapore General Hospital, Singapore, Singapore. [8]SingHealth AI Office, Singapore Health Services, Singapore, Singapore. [9]Institute of Data Science, National University of Singapore, Singapore, Singapore.

Received: 24 May 2022 Accepted: 25 October 2022
Published online: 04 November 2022

References

1. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338:b375.
2. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
3. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules - Applications and methodological standards. N Engl J Med. 1985;313(13):793–9.
4. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. Am Heart J. 1991;121(1 Pt 2):293–8.
5. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. Ann Emerg Med. 1992;21(4):384–90.
6. Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, et al. A prognostic index in primary breast cancer. Br J Cancer. 1982;45(3):361–6.
7. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989;81(24):1879–86.
8. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardiothorac Surg. 1999;16(1):9–13.
9. Stenhouse C, Coates S, Tivey M, Allsop P, Parker T. Prospective evaluation of a modified early warning score to aid earlier detection of patients developing critical illness on a general surgical ward. Br J Anaesth. 2000;84(5):663P.
10. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified early warning score in medical admissions. QJM. 2001;94(10):521–6.
11. Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, et al. A simplified acute physiology score for ICU patients. Crit Care Med. 1984;12(11):975–7.
12. Wang LE, Shaw PA, Mathelier HM, Kimmel SE, French B. Evaluating risk-prediction models using data from electronic health records. Ann Appl Stat. 2016;10(1):286–304.
13. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144–51.
14. Heinze G, Wallisch C, Dunkler D. Variable selection - a review and recommendations for the practicing statistician. Biom J. 2018;60(3):431–49.
15. Xie F, Chakraborty B, Ong MEH, Goldstein BA, Liu N. AutoScore: a machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records. JMIR Med Inform. 2020;8(10):e21798.
16. Xie F, Ong MEH, Liew JNMH, Tan KBK, Ho AFW, Nadarajan GD, et al. Development and assessment of an interpretable machine learning triage tool for estimating mortality after emergency admissions. JAMA Netw Open. 2021;4(8):e2118467.
17. Wong XY, Ang YK, Li K, Chin YH, Lam SSW, Tan KBK, et al. Development and validation of the SARICA score to predict survival after return of spontaneous circulation in out of hospital cardiac arrest using an interpretable machine learning framework. Resuscitation. 2022;170:126–33.
18. Petersen KK, Lipton RB, Grober E, Davatzikos C, Sperling RA, Ezzati A. Predicting amyloid positivity in cognitively unimpaired older adults. Neurology. 2022;98(24):e2425–35.
19. Liu N, Liu M, Chen X, Ning Y, Lee JW, Siddiqui FJ, et al. Development and validation of an interpretable prehospital return of spontaneous circulation (P-ROSC) score for patients with out-of-hospital cardiac arrest using machine learning: a retrospective study. eClinicalMedicine. 2022;48:101422.
20. Ang Y, Li S, Ong MEH, Xie F, Teo SH, Choong L, et al. Development and validation of an interpretable clinical score for early identification of acute kidney injury at the emergency department. Sci Rep. 2022;12(1):1–8.
21. Kanagarathinam K, Sankaran D, Manikandan R. Machine learning-based risk prediction model for cardiovascular disease using a hybrid dataset. Data Knowl Eng. 2022;140:102042.
22. Zhao Y, Li X, Li S, Dong M, Yu H, Zhang M, et al. Using machine learning techniques to develop risk prediction models for the risk of incident diabetic retinopathy among patients with type 2 diabetes mellitus: a cohort study. Front Endocrinol (Lausanne). 2022;13:885.
23. Adi NS, Farhany R, Ghina R, Napitupulu H. Stroke Risk Prediction Model Using Machine Learning. In: 2021 International Conference on Artificial Intelligence and Big Data Analytics; 2021. p. 56–60.
24. Li X, Wang Y, Xu J. Development of a machine learning-based risk prediction model for cerebral infarction and comparison with nomogram model. J Affect Disord. 2022;314:341–8.
25. Pera M, Gibert J, Gimeno M, Garsot E, Eizaguirre E, Miró M, et al. Machine learning risk prediction model of 90-day mortality after gastrectomy for Cancer. Ann Surg. 2022;276:776–83.
26. Jiang H, Mao H, Lu H, Lin P, Garry W, Lu H, et al. Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease. Int J Med Inform. 2021;145:104326.
27. Kawakami E, Tabata J, Yanaihara N, Ishikawa T, Koseki K, Iida Y, et al. Application of artificial intelligence for preoperative diagnostic and prognostic prediction in epithelial ovarian cancer based on blood biomarkers. Clin Cancer Res. 2019;25(10):3006–15.
28. Valenta Z, Pitha J, Poledne R. Proportional odds logistic regression--effective means of dealing with limited uncertainty in dichotomizing clinical outcomes. Stat Med. 2006;25(24):4227–34.
29. Roozenbeek B, Lingsma HF, Perel P, Edwards P, Roberts I, Murray GD, et al. The added value of ordinal analysis in clinical trials: an example in traumatic brain injury. Crit Care. 2011;15(3):R127.
30. McHugh GS, Butcher I, Steyerberg EW, Marmarou A, Lu J, Lingsma HF, et al. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT project. Clin Trials. 2010;7(1):44–57.
31. Saver JL. Novel end point analytic techniques and interpreting shifts across the entire range of outcome scales in acute stroke trials. Stroke. 2007;38(11):3055–62.
32. Machado SG, Murray GD, Teasdale GM. Evaluation of designs for clinical trials of neuroprotective agents in head injury. European Brain Injury Consortium. J Neurotrauma. 1999;16(12):1131–8.
33. Ceyisakar IE, van Leeuwen N, Dippel DW, Steyerberg EW, Lingsma HF. Ordinal outcome analysis improves the detection of between-hospital differences in outcome. BMC Med Res Methodol. 2021;21(4):4.
34. Uryniak T, Chan ISF, Fedorov VV, Jiang Q, Oppenheimer L, Snapinn SM, et al. Responder analyses—a PhRMA position paper. Stat Biopharm Res. 2011;3(3):476–87.

35. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006;332(7549):1080.
36. Lingsma HF, Bottle A, Middleton S, Kievit J, Steyerberg EW, Marang-van de Mheen PJ. Evaluation of hospital outcomes: the relation between length-of-stay, readmission, and mortality in a large international administrative database. BMC Health Serv Res. 2018;18(1):116.
37. Myers J, Kei J, Aithal S, Aithal V, Driscoll C, Khan A, et al. Diagnosing middle ear dysfunction in 10- to 16-month-old infants using wideband absorbance: an ordinal prediction model. J Speech Lang Hear Res. 2019;62(8):2906–17.
38. Edlinger M, Dörler J, Ulmer H, Wanitschek M, Steyerberg EW, Alber HF, et al. An ordinal prediction model of the diagnosis of non-obstructive coronary artery and multi-vessel disease in the CARDIIGAN cohort. Int J Cardiol. 2018;267:8–12.
39. Sawhney R, Joshi H, Gandhi S, Jin D, Shah RR. Robust suicide risk assessment on social media via deep adversarial learning. J Am Med Inform Assoc. 2021;28(7):1497–506.
40. Barbero-Gómez J, Gutiérrez PA, Vargas VM, Vallejo-Casas JA, Hervás-Martínez C. An ordinal CNN approach for the assessment of neurological damage in Parkinson's disease patients. Expert Syst Appl. 2021;182:115271.
41. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15.
42. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
43. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London: Chapman and Hall/CRC; 1989.
44. McCullagh P. Regression models for ordinal data. J R Stat Soc Ser B. 1980;42(2):109–42.
45. Rosati R, Romeo L, Vargas VM, Gutiérrez PA, Hervás-Martínez C, Frontoni E. A novel deep ordinal classification approach for aesthetic quality control classification. Neural Comput Applic. 2022;34(14):11625–39.
46. Wang L, Zhu D. Tackling ordinal regression problem for heterogeneous data: sparse and deep multi-task learning approaches. Data Min Knowl Disc. 2021;35(3):1134.
47. van Calster B, van Belle V, Vergouwe Y, Steyerberg EW. Discrimination ability of prediction models for ordinal outcomes: relationships between existing measures and a new measure. Biom J. 2012;54(5):674–85.
48. Waegeman W, de Baets B, Boullart L. ROC analysis in ordinal regression learning. Pattern Recogn Lett. 2008;29(1):1–9.
49. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982;247(18):2543–6.
50. Harrell FEJ. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. New York: Springer; 2015. (Springer Series in Statistics)
51. DiCiccio TJ, Efron B. Bootstrap confidence intervals. Stat Sci. 1996;11(3):189–228.
52. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies. Int J Med Inform. 2021;153:104510.
53. Xie F, Liu N, Wu SX, Ang Y, Low LL, Ho AFW, et al. Novel model for predicting inpatient mortality after emergency admission to hospital in Singapore: retrospective observational study. BMJ Open. 2019;9(9):e031382.
54. Liu N, Xie F, Siddiqui FJ, Wah Ho AF, Chakraborty B, Nadarajan GD, et al. Leveraging Large-Scale Electronic Health Records and Interpretable Machine Learning for Clinical Decision Making at the Emergency Department: Protocol for System Development and Validation. JMIR Res Protoc. 2022;11(3):e34201.
55. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2020. Available from: https://cran.r-project.org
56. Christensen RHB. ordinal---Regression Models for Ordinal Data. R package version 2018.4–19. 2018. Available from: http://www.cran.r-project.org/package=ordinal/
57. Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York: Springer; 2002.
58. Wurm MJ, Rathouz PJ, Hanlon BM. Regularized ordinal regression and the ordinalNet R package. Journal of Statistical Software. 2017;99(6):1–42.
59. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2(3):18–22.
60. Kropko J, Harden JJ. coxed: Duration-Based Quantities of Interest for the Cox Proportional Hazards Model; 2020. Available from: https://CRAN.R-project.org/package=coxed.
61. Harrell Jr F. Hmisc: Harrell Miscellaneous; 2021. Available from: https://CRAN.R-project.org/package=Hmisc.
62. Goff DCJ, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association task force on practice guidelines. Circulation. 2014;129(25 Suppl 2):S49–73.
63. Rabar S, Lau R, O'Flynn N, Li L, Barry P. Risk assessment of fragility fractures: summary of NICE guidance. BMJ. 2012;345:e3698.
64. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
65. Probst P, Boulesteix A-L. To tune or not to tune the number of trees in random Forest. J Mach Learn Res. 2018;18:1–18.
66. Scurfield BK. Multiple-event forced-choice tasks in the theory of signal detectability. J Math Psychol. 1996;40(3):253–69.
67. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn. 2001;45(2):171–86.
68. van Calster B, van Belle V, Vergouwe Y, Timmerman D, van Huffel S, Steyerberg EW. Extending the c-statistic to nominal polytomous outcomes: the Polytomous discrimination index. Stat Med. 2012;31(23):2610–26.
69. Dover DC, Islam S, Westerhout CM, Moore LE, Kaul P, Savu A. Computing the polytomous discrimination index. Stat Med. 2021;40(16):3667–81.
70. Guo X, Khalid MA, Domingos I, Michala AL, Adriko M, Rowel C, et al. Smartphone-based DNA diagnostics for malaria detection using deep learning for local decision support and blockchain technology for security. Nat Electron. 2021;4(8):615–24.
71. Krittanawong C, Rogers AJ, Johnson KW, Wang Z, Turakhia MP, Halperin JL, et al. Integration of novel monitoring devices with machine learning technology for scalable cardiovascular management. Nat Rev Cardiol. 2020;18(2):75–91.
72. Wu Y, Yao X, Vespasiani G, Nicolucci A, Dong Y, Kwong J, et al. Mobile app-based interventions to support diabetes self-management: a systematic review of randomized controlled trials to identify functions associated with glycemic efficacy. JMIR Mhealth Uhealth. 2017;5(3):e6522.
73. Ferri A, Rosati R, Bernardini M, Gabrielli L, Casaccia S, Romeo L, et al. Towards the Design of a Machine Learning-based Consumer Healthcare Platform powered by Electronic Health Records and measurement of Lifestyle through Smartphone Data. In: 2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT); 2019. p. 37–40.

Publisher's Note