RESEARCH ARTICLE

A novel interpretable machine learning system to generate clinical risk scores: An application for predicting early mortality or unplanned readmission in a retrospective cohort study

Yilin Ning[1], **Siqi Li**[1], **Marcus Eng Hock Ong**[2,3,4], **Feng Xie**[1,2], **Bibhas Chakraborty**[1,2,5,6], **Daniel Shu Wei Ting**[1,7,8], **Nan Liu**[1,2,3,8,9]*

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** Health Services Research Centre, Singapore Health Services, Singapore, Singapore, **4** Department of Emergency Medicine, Singapore General Hospital, Singapore, Singapore, **5** Department of Statistics and Data Science, National University of Singapore, Singapore, Singapore, **6** Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, United States of America, **7** Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore, **8** SingHealth AI Health Program, Singapore Health Services, Singapore, Singapore, **9** Institute of Data Science, National University of Singapore, Singapore, Singapore

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

Abstract

Risk scores are widely used for clinical decision making and commonly generated from logistic regression models. Machine-learning-based methods may work well for identifying important predictors to create parsimonious scores, but such 'black box' variable selection limits interpretability, and variable importance evaluated from a single model can be biased. We propose a robust and interpretable variable selection approach using the recently developed Shapley variable importance cloud (ShapleyVIC) that accounts for variability in variable importance across models. Our approach evaluates and visualizes overall variable contributions for in-depth inference and transparent variable selection, and filters out non-significant contributors to simplify model building steps. We derive an ensemble variable ranking from variable contributions across models, which is easily integrated with an automated and modularized risk score generator, AutoScore, for convenient implementation. In a study of early death or unplanned readmission after hospital discharge, ShapleyVIC selected 6 variables from 41 candidates to create a well-performing risk score, which had similar performance to a 16-variable model from machine-learning-based ranking. Our work contributes to the recent emphasis on interpretability of prediction models for high-stakes decision making, providing a disciplined solution to detailed assessment of variable importance and transparent development of parsimonious clinical risk scores.

Citation: Ning Y, Li S, Ong MEH, Xie F, Chakraborty B, Ting DSW, et al. (2022) A novel interpretable machine learning system to generate clinical risk scores: An application for predicting early mortality or unplanned readmission in a retrospective cohort study. PLOS Digit Health 1(6): e0000062. https://doi.org/10.1371/journal.pdig.0000062

Editor: Henry Horng-Shing Lu, National Yang Ming Chiao Tung University, TAIWAN

Received: February 8, 2022

Accepted: May 10, 2022

Published: June 13, 2022

Data Availability Statement: Due to ethical reasons and institutional guidelines, the data presented in the study cannot be shared publicly. Data are available to researchers with some access restrictions applied upon request. Interested researchers may contact SingHealth Health Services Research Centre (Email: hsr@singhealth.com.sg) for more details.

Funding: Yilin Ning 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 funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Risk scores help clinicians quickly assess the risk for a patient by adding up a few scores associated with key predictors. Given the simplicity of such scores, shortlisting the most important predictors is key to predictive performance, but traditional methods are sometimes insufficient when there are a lot of candidates to choose from. As a rising area of research, machine learning provides a growing toolkit for variable selection, but as many machine learning models are complex 'black boxes' that differ considerably from risk scores, directly plugging machine learning tools into risk score development can harm both interpretability and predictive performance. We propose a robust and interpretable variable selection mechanism that is tailored to risk scores, and integrate it with an automated framework for convenient risk score development. In a clinical example, we demonstrated how our proposed method can help researchers understand the contribution of 41 candidate variables to outcome prediction through visualizations, filter out 20 variables with non-significant contribution and build a well-performing risk score using only 6 variables, whereas a machine-learning-based method selected 16 variables to achieve a similar performance. We have thus presented a useful tool to support transparent high-stakes decision making.

Introduction

Scoring models are widely used in clinical settings for assessment of risks and guiding decision making [1]. For example, the LACE index [2] for predicting unplanned readmission and early death after hospital discharge has been applied and validated in various clinical and public health settings since it was developed in 2010 [3–5]. The LACE index is appreciated for its simplicity, achieving moderate discrimination power [6] by using only four basic components: inpatient length of stay (LOS), acute admission, comorbidity, and the number of emergency department (ED) visits in past 6 months. There have been ongoing efforts to derive new readmission risk prediction models for improved performance and for specific subcohorts, which, despite the increasing availability of machine learning methods in recent years, are dominated by the traditional logistic regression approach [7,8].

Logistic regression models generate transparent prediction models that are easily converted to risk scores, e.g., by rounding the regression coefficients [9], and model interpretability is mainly reflected by variable selection. Traditional variable selection approaches (e.g., stepwise selection and penalized likelihood) are straightforward, but there have been methodological discussions on their insufficiency in identifying important predictors and controlling model complexity [10,11]. Machine learning methods are well-performing alternatives, e.g., a study used decision trees and neural networks to heuristically develop a 11-variable logistic regression model from 76 candidate variables [12], and a recently developed automated risk score generator, the AutoScore [13], uses the random forest (RF) to rank variables and built well-performing scoring models using less than 10 variables in clinical applications [13,14]. Such machine learning methods rank variables by their importance to best-performing model(s) trained from data, but the 'black box' nature of most machine learning models (e.g., neural networks and tree ensembles such as the RF and XGBoost [15]) limits the interpretability of the variable selection steps. The interpretability issue may be alleviated via post-hoc explanation, for which the widely used Shapley additive explanations (SHAP) [16] is the current state-of-the-art. However, another general issue in current variable importance analyses has been largely ignored: restricting the assessment to best-performing models can lead to biased perception on variable importance [17].

Since predictive performance is often not the only concern in practice, a recent study extended the investigation to a group of models that are 'good enough' to generate a variable importance cloud (VIC), which provides a comprehensive overview of how important (or unimportant) each variable could be to accurate prediction [17]. Our recent work combined VIC with the well-received Shapley values to devise a Shapley variable importance cloud (ShapleyVIC) [18], which easily integrates with current practice to provide less biased assessments of variable importance. In a recidivism risk prediction study, SHAP analysis of the optimal model reported high importance for race, whereas the low overall importance estimated by both ShapleyVIC and VIC suggested that the assessment based on the single model was likely an overclaim [18]. Moreover, ShapleyVIC reports uncertainty intervals for overall variable importance for statistical assessments, which is not easily available from VIC [17,18]. In a mortality risk prediction study using logistic regression [18], ShapleyVIC handled strong collinearity among variables and effectively communicated findings through statistics and visualizations. These applications suggest ShapleyVIC as a suitable method for guiding variable selection steps when developing clinical scores from regression models.

In this paper, we explore the use of ShapleyVIC as an interpretable and robust variable selection approach for developing clinical scores. Specifically, we use an electronic health record dataset from ED to predict death or unplanned readmission within 30 days after hospital discharge. We interpret the importance of candidate variables with respect to a group of 'good' logistic regression models using visualizations of ShapleyVIC values, combine information across models to generate an ensemble variable ranking, and use the estimated overall importance to filter out candidate variables that will likely add noise to the scoring model. To develop scoring models from ShapleyVIC-ranked candidate variables, we take advantage of the disciplined and modularized AutoScore framework, which grows a relatively simple, interpretable model based on a ranked list of variables until the inclusion of any additional variable makes little improvement to predictive performance [13]. The current AutoScore framework ranks variables using the RF, which is a commonly used machine learning method in clinical applications that performs well and is easy to train. The inability of traditional variable selection approaches (i.e., stepwise variable selection and penalized likelihood approach) in building sparse prediction models and the good performance of the RF-based approach has been previously demonstrated [13]. Hence, in this work we include RF-based scoring models as a baseline to evaluate the ShapleyVIC-based model and to demonstrate the benefits of ShapleyVIC that are not available from existing variable importance methods. We also compare these scoring models to the LACE index to show improvements over existing models.

Results

Study cohort

This study aimed to develop a scoring model to predict unplanned readmission or death within 30 days after hospital discharge. The data was derived from a retrospective cohort study of cases who visited the ED of Singapore General Hospital and were subsequently admitted to the hospital. The full cohort consists of 411,137 cases, where 388,576 cases were eligible. 63,938 (16.5%) eligible cases developed the outcome of interest, where 8174 were death and 55,764 had readmission. As summarized in Table 1, cases with and without the outcome were significantly different (i.e., had p-value<0.05) in all 41 candidate variables except the use of ventilation. Specifically, compared to those without the outcome, cases with outcome tended to be older and have shorter ED LOS, higher ED triage, longer inpatient LOS, more ED visits in the

Table 1. Descriptive statistics of the full cohort. The outcome is death or readmission within 30 days after hospital discharge.

	Overall (N = 388,576)	With outcome (N = 63,938, 16.5%)	Without outcome (N = 324,638, 83.5%)	p-value**
Patient demographics				
Age (years): mean (SD)	62.0 (17.8)	67.6 (15.3)	60.9 (18.0)	<0.001
Gender: n (%)				<0.001
Female	195426 (50.3)	30405 (47.6)	165021 (50.8)	
Male	193150 (49.7)	33533 (52.4)	159617 (49.2)	
Race: n (%)				<0.001
Chinese	274929 (70.8)	48079 (75.2)	226850 (69.9)	
Indian	41718 (10.7)	6277 (9.8)	35441 (10.9)	
Malay	47528 (12.2)	7430 (11.6)	40098 (12.4)	
Others	24401 (6.3)	2152 (3.4)	22249 (6.9)	
ED admission				
ED LOS (hours): mean (SD)	2.8 (1.7)	2.5 (1.6)	2.8 (1.7)	<0.001
ED triage: n (%)				<0.001
P1	69383 (17.9)	14751 (23.1)	54632 (16.8)	
P2	219245 (56.4)	39287 (61.4)	179958 (55.4)	
P3 and P4	99948 (25.7)	9900 (15.5)	90048 (27.7)	
ED boarding time (hours): mean (SD)	4.8 (3.7)	4.8 (3.9)	4.8 (3.7)	0.014
Consultation waiting time (hours): mean (SD)	0.8 (0.8)	0.7 (0.7)	0.8 (0.8)	<0.001
Day of week: n (%)				<0.001
Friday	54461 (14.0)	8969 (14.0)	45492 (14.0)	
Monday	64914 (16.7)	10549 (16.5)	54365 (16.7)	
Weekend	99797 (25.7)	16903 (26.4)	82894 (25.5)	
Midweek	169404 (43.6)	27517 (43.0)	141887 (43.7)	
Inpatient admission				
Inpatient LOS (days): mean (SD)	6.54 (11.66)	8.19 (12.17)	6.21 (11.52)	<0.001
Admission type				<0.001
A1	14964 (3.9)	1401 (2.2)	13563 (4.2)	
B1	32688 (8.4)	3144 (4.9)	29544 (9.1)	
B2	185244 (47.7)	27713 (43.3)	157531 (48.5)	
C	155680 (40.1)	31680 (49.5)	124000 (38.2)	
Healthcare utilization: mean (SD)				
No. ED visits in past 6 months	0.59 (1.42)	1.58 (2.59)	0.39 (0.93)	<0.001
No. surgery in past year	0.20 (0.75)	0.43 (1.10)	0.16 (0.65)	<0.001
No. ICU stays in past year	0.02 (0.26)	0.05 (0.36)	0.02 (0.23)	<0.001
No. HD stays in past year	0.09 (0.47)	0.17 (0.69)	0.07 (0.40)	<0.001
Vital sign and clinical tests at ED				
Ventilation: n (%)	53 (0.0)	8 (0.0)	45 (0.0)	0.789
Resuscitation: n (%)	4136 (1.1)	874 (1.4)	3262 (1.0)	<0.001
Pulse, beat/minute: mean (SD)*	82.33 (16.81)	84.56 (17.51)	81.89 (16.63)	<0.001
Respiration, breath/minute: mean (SD)*	17.86 (1.65)	18.08 (1.90)	17.81 (1.59)	<0.001
SpO_2, %: mean (SD)*	98.01 (3.05)	97.87 (3.28)	98.04 (3.00)	<0.001
DBP, mmHg: mean (SD)*	71.56 (13.36)	70.60 (13.66)	71.75 (13.29)	<0.001
SBP, mmHg: mean (SD)*	134.46 (25.32)	134.23 (26.19)	134.50 (25.14)	0.013
Bicarbonate, mmol/L: mean (SD)*	22.80 (3.65)	22.50 (4.16)	22.86 (3.54)	<0.001
Creatinine, μmol/L: mean (SD)*	153.28 (207.43)	203.71 (249.54)	143.11 (196.30)	<0.001
Potassium, mmol/L: mean (SD)*	4.16 (0.71)	4.25 (0.78)	4.14 (0.69)	<0.001

(Continued)

Table 1. (Continued)

	Overall (N = 388,576)	With outcome (N = 63,938, 16.5%)	Without outcome (N = 324,638, 83.5%)	p-value**
Sodium, mmol/L: mean (SD)*	135.07 (5.02)	134.13 (5.77)	135.26 (4.83)	<0.001
Comorbidity: n (%)				
Myocardial infarction	22064 (5.7)	7408 (11.6)	14656 (4.5)	<0.001
Congestive heart failure	43758 (11.3)	13357 (20.9)	30401 (9.4)	<0.001
Peripheral vascular disease	22239 (5.7)	6883 (10.8)	15356 (4.7)	<0.001
Stroke	50400 (13)	11719 (18.3)	38681 (11.9)	<0.001
Dementia	11178 (2.9)	3291 (5.1)	7887 (2.4)	<0.001
Pulmonary	37652 (9.7)	9722 (15.2)	27930 (8.6)	<0.001
Rheumatic	5507 (1.4)	1213 (1.9)	4294 (1.3)	<0.001
Peptic ulcer disease	14678 (3.8)	3888 (6.1)	10790 (3.3)	<0.001
Mild liver disease	18402 (4.7)	4857 (7.6)	13545 (4.2)	<0.001
Severe liver disease	5893 (1.5)	2296 (3.6)	3597 (1.1)	<0.001
Diabetes (without complications)	49057 (12.6)	9918 (15.5)	39139 (12.1)	<0.001
Diabetes with complications	96334 (24.8)	22601 (35.3)	73733 (22.7)	<0.001
Paralysis	21563 (5.5)	5248 (8.2)	16315 (5)	<0.001
Renal	80895 (20.8)	22566 (35.3)	58329 (18)	<0.001
Cancer (non-metastatic)	34385 (8.8)	8532 (13.3)	25853 (8)	<0.001
Metastatic cancer	28630 (7.4)	11049 (17.3)	17581 (5.4)	<0.001
CCI: mean (SD)	2.41 (2.68)	4.14 (3.00)	2.07 (2.48)	<0.001

*: Excluding 7521 missing entries for pulse, 8605 missing entries for respiration, 8533 missing entries for SpO_2, 4175 missing entries for DBP, 4181 missing entries for and SBP, 48,739 missing entries for bicarbonate, 48,695 missing entries for creatinine, 50,358 missing entries for potassium, and 48,637 missing entries for sodium.

**: P-values were reported from two-sample t-tests for continuous variables and Chi-squared tests for categorical variables.

CCI: Charlson comorbidity index; 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.

https://doi.org/10.1371/journal.pdig.0000062.t001

past 6 months and more comorbid. The training, validation and test sets consisted of 272,004 (70%), 38,857 (10%) and 77,715 (20%) cases randomly drawn from the full cohort, respectively.

RF-generated models

We first describe scoring models generated using RF-based variable ranking, which ranked the 41 variables based on their importance to the RF trained on the training set and used it in the variable ranking module of the AutoScore framework. Fig 1 presents the parsimony plot from AutoScore, which visualizes the improvement in model performance (measured using the area under the receiver operating characteristic curve [AUC] on the validation set) when each variable entered the scoring model in descending order of importance. The AutoScore guideline suggested two feasible models: Model 1A using the first 6 variables (i.e., number of ED visits in past 6 months, inpatient LOS, ED LOS, ED boarding time, creatinine, and age) where the inclusion of the next variable only improved AUC by 0.2%, and Model 1B using the first 16 variables to include the 16th variable (i.e., metastatic cancer) that resulted in a pronounced increase (1.8%) in AUC. The 6-variable Model 1A had an AUC of 0.739 (95% confidence interval [CI]: 0.734–0.743; see Table 2) evaluated on the test set, comparable to that of the LACE index (AUC = 0.733, 95% CI: 0.728–0.738), whereas Model 1B outperformed the LACE index with an AUC of 0.759 (95% CI: 0.754–0.764). Researchers may predict the outcome for a

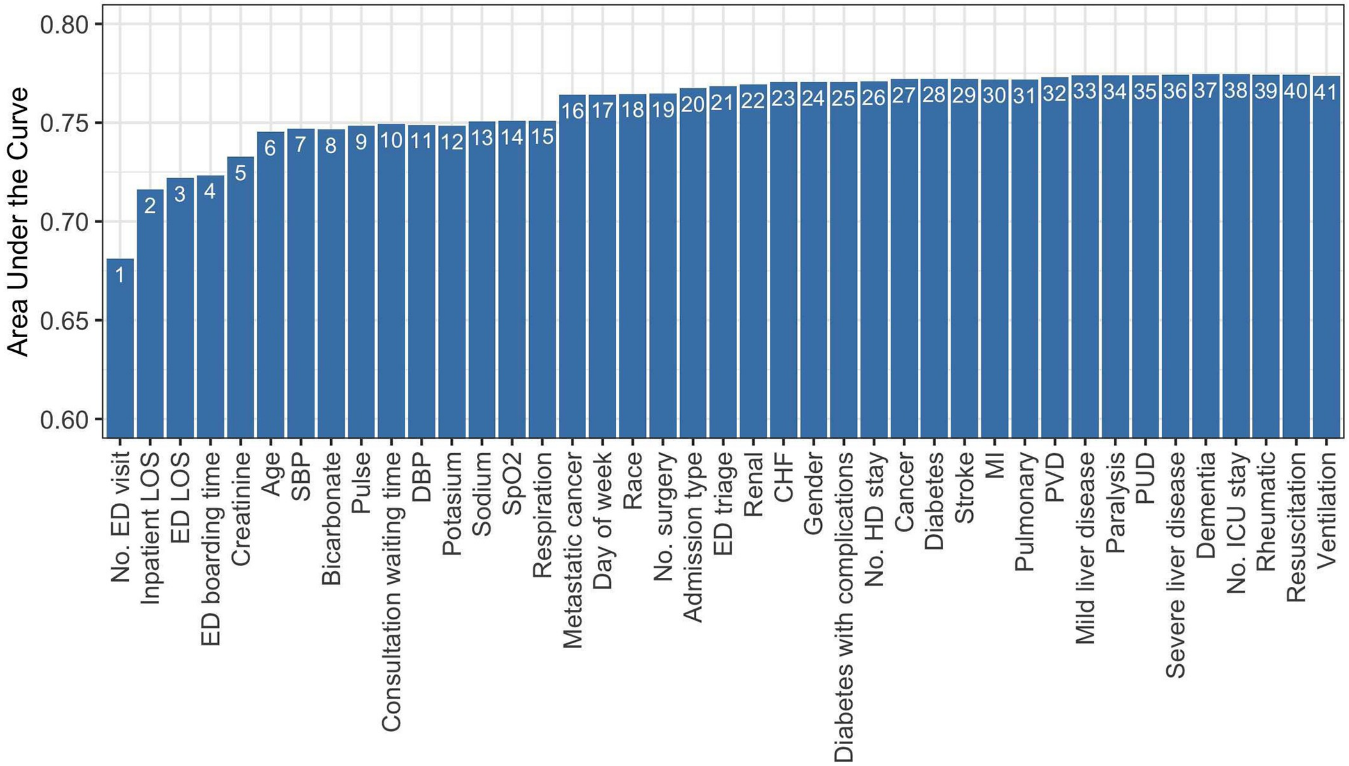

Fig 1. Parsimony plot on validation set using random-forest-based variable ranking. Number of ED visit is within 6 months before current inpatient stay. Number of surgery, ICU stay and HD stay are within 1 year before current inpatient stay. CHF: Congestive heart failure; DBP: diastolic blood pressure; ED: emergency department; HD: high dependency ward; ICU: intensive care unit; LOS: length of stay; MI: Myocardial infarction; PVD: Peripheral vascular disease; PUD: Peptic ulcer disease; SBP: systolic blood pressure; SpO_2: blood oxygen saturation.

https://doi.org/10.1371/journal.pdig.0000062.g001

patient by applying a threshold to the derived risk score, for example, by using the optimal threshold defined as the point nearest to the upper-left corner of the receiver operating characteristic (ROC) curve (for which additional performance metrics are reported in Table 2), or by manually selecting a threshold that reaches a desirable level of sensitivity and/or specificity. Table 3 presents the scoring tables for Models 1A and 1B after fine-tuning the cut-off values

Table 2. Model performance evaluated on the test set.

Method	Model	Number of variables	AUC (95% CI)	Accuracy** (95% CI)	Sensitivity** (95% CI)	Specificity** (95% CI)	F1-score** (95% CI)
LACE	--	3+17*	0.733 (0.728, 0.738)	0.641 (0.638, 0.645)	0.735 (0.727, 0.743)	0.623 (0.619, 0.627)	0.788 (0.744, 0.793)
AutoScore	Model 1A	6	0.739 (0.734, 0.743)	0.691 (0.687, 0.694)	0.654 (0.646, 0.663)	0.698 (0.694, 0.701)	0.790 (0.788, 0.793)
	Model 1B	16	0.759 (0.754, 0.764)	0.685 (0.682, 0.688)	0.704 (0.696, 0.712)	0.681 (0.677, 0.684)	0.788 (0.783, 0.793)
AutoScore-ShapleyVIC	Model 2	6	0.756 (0.751, 0.760)	0.710 (0.707, 0.713)	0.660 (0.652, 0.668)	0.720 (0.716, 0.723)	0.793 (0.788, 0.806)

*: LACE is computed from inpatient length of stay, number of ED visits in past 6 months, acute admission and Charlson comorbidity index (which is computed from 17 comorbidities). See Table A in S1 Text for detailed scoring table.

**: Statistics are evaluated at optimal score thresholds (defined as the point nearest to the upper-left corner of the receiver operating characteristic curve) for each model: 10 for LACE, 42 for Model 1A, 33 for Model 1B and 31 for Model 2.

https://doi.org/10.1371/journal.pdig.0000062.t002

Table 3. Scoring tables of Models 1A and 1B generated using AutoScore, with variables ranked by random forest.

Variable	Interval	Point	
		Model 1A	Model 1B
Number of ED visits in past 6 months	<1	0	0
	[1,3)	17	11
	>=3	40	27
Inpatient LOS (days)	<1	0	0
	[1,2)	1	1
	[2,7)	6	3
	>=7	12	6
ED LOS	<40min	12	5
	[40min, 80min)	9	3
	[80min, 4h)	6	2
	[4h, 6h)	2	1
	>=6h	0	0
ED boarding time	<80min	0	0
	[80min, 6.5h)	2	1
	>=6.5h	3	1
Creatinine (μmol/L)	<45	7	2
	[45, 60)	1	0
	[60, 135)	0	1
	[135, 600)	7	5
	>=600	8	7
Age (years)	<25	0	0
	[25,45)	7	4
	[45,75)	18	10
	[75,85)	21	13
	>=85	25	16
Systolic blood pressure (mmHg)	<100	--	3
	[100, 115)	--	2
	[115, 155)	--	1
	>=155	--	0
Bicarbonate (mmol/L)	<17	--	1
	[17, 20)	--	1
	[20, 28)	--	0
	>=28	--	3
Pulse	<60	--	0
	[60, 70)	--	1
	[70, 100)	--	2
	>=100	--	3
Consultation waiting time (hours)	<1.5	--	2
	[1.5, 2.5)	--	1
	>=2.5	--	0
Diastolic blood pressure (mmHg)	<50	--	1
	[50, 95)	--	0
	>=95	--	1
Potassium (mmol/L)	<3.5	--	2
	[3.5, 4)	--	0
	[4, 4.5)	--	1
	>=4.5	--	2

(*Continued*)

Table 3. (Continued)

Variable	Interval	Point	
		Model 1A	Model 1B
Sodium (mmol/L)	<125	--	6
	[125,130)	--	4
	[130,135)	--	2
	>=135	--	0
SpO_2	<95	--	1
	>=95	--	0
Respiration	<17	--	0
	[17, 20)	--	1
	>=20	--	3
Metastatic cancer	No	--	0
	Yes	--	15

"[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 stay.

https://doi.org/10.1371/journal.pdig.0000062.t003

for continuous variables, and the scoring table of the LACE index is provided as Table A in S1 Text for readers' convenience.

ShapleyVIC-generated model

Next, we describe the assessment of overall variable importance using ShapleyVIC values and the implications, the resulting ensemble ranking that accounts for variabilities in variable importance across models, the simplification of model building steps based on the significance of overall variable importance, and the final scoring model developed. Using the training set, we trained a logistic regression model on all 41 candidate predictors by minimizing model loss. Defining "nearly optimal" as model loss exceeding the minimum loss by no more than 5%, we randomly generated 350 models from the corresponding model space to empirically study variable importance across nearly optimal models. We used the first 3500 cases in the validation set to evaluate the ShapleyVIC values of these 350 models, which we found enough for the algorithm to converge and generate stable values in preliminary experiments.

Unlike machine-learning-based methods (e.g., RF as described in the previous subsection) that focus on relative importance of candidate variables for ranking purpose, ShapleyVIC quantifies the extent of variable importance for more in-depth inference, and communicates the findings through different forms of visualizations to facilitate interpretation. Fig 2 visualizes the average ShapleyVIC values across 350 models and the 95% prediction intervals (PIs), which estimate overall variable importance and the variability, respectively. Non-positive values indicate unimportance and are visualized using grey bars. Fig 3 presents the distribution of ShapleyVIC values from individual models to visualize the relationship between variable importance and model performance among the nearly optimal models. While both the RF (see Fig 1) and the average ShapleyVIC values (see Fig 2) suggested the number of previous ED visits as the most important variable among the 41 candidates, the 95% PI of the average ShapleyVIC value (see Fig 2) further concluded the significance of its overall importance, and the violin plot in Fig 3 revealed its much higher importance than other variables in all nearly optimal models studied. Metastatic cancer had the second highest average ShapleyVIC value

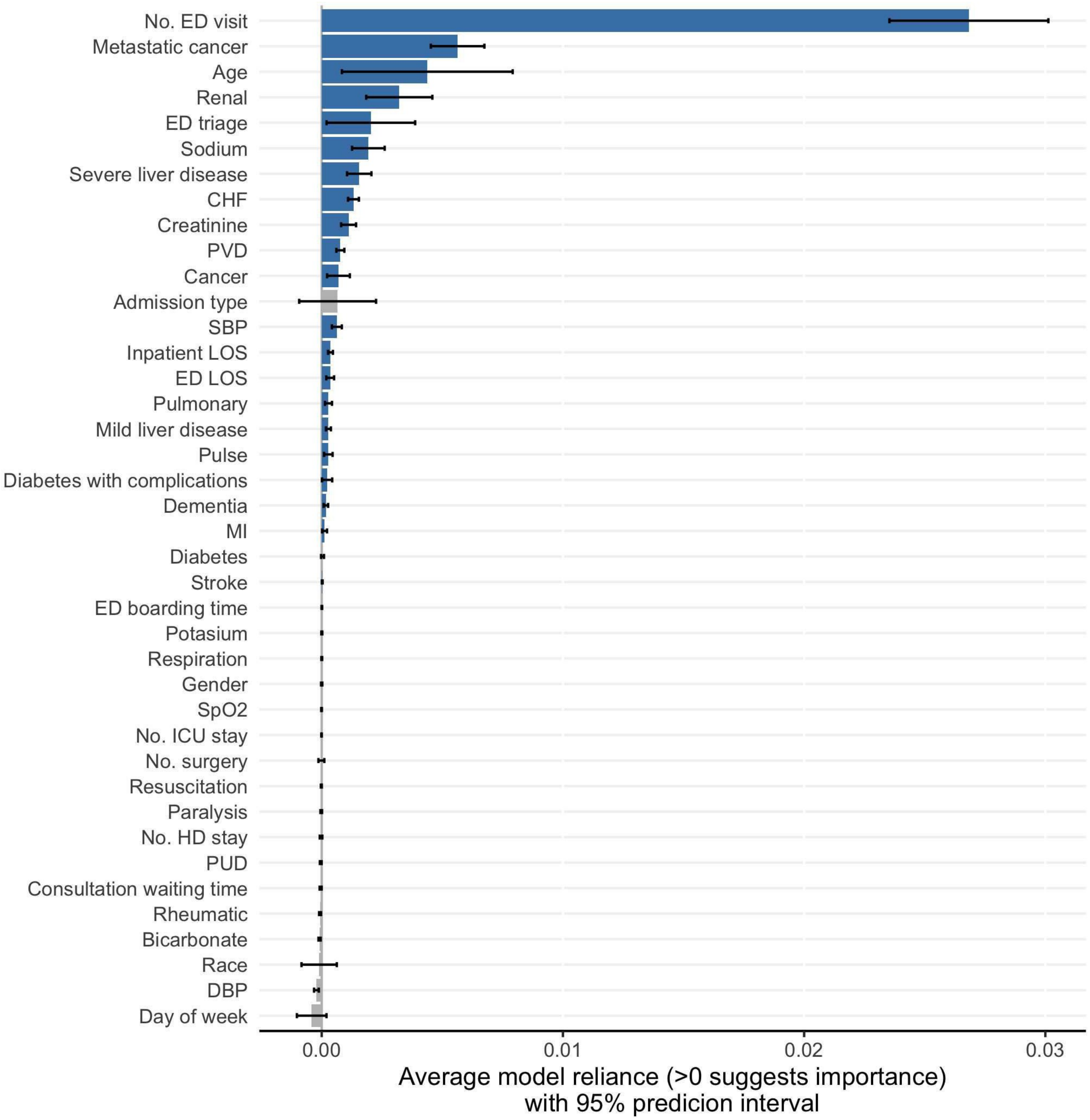

Fig 2. Average ShapleyVIC values and 95% prediction intervals from 350 nearly optimal logistic regression models. Ventilation had zero ShapleyVIC value in all 350 models analyzed, resulting in a 95% prediction interval centered at zero with zero length and was therefore not plotted. Number of ED visit is within 6 months before current inpatient stay. Number of surgery, ICU stay and HD stay are within 1 year before current inpatient stay. CHF: Congestive heart failure; DBP: diastolic blood pressure; ED: emergency department; HD: high dependency ward; ICU: intensive care unit; LOS: length of stay; MI: Myocardial infarction; PVD: Peripheral vascular disease; PUD: Peptic ulcer disease; SBP: systolic blood pressure; SpO_2: blood oxygen saturation.

https://doi.org/10.1371/journal.pdig.0000062.g002

among all variables, consistent with its considerable contribution to model performance observed from the RF-based variable ranking.

Both metastatic cancer and age were important to all nearly optimal models studied (indicated by positive ShapleyVIC values), but the wider spread of ShapleyVIC values for age

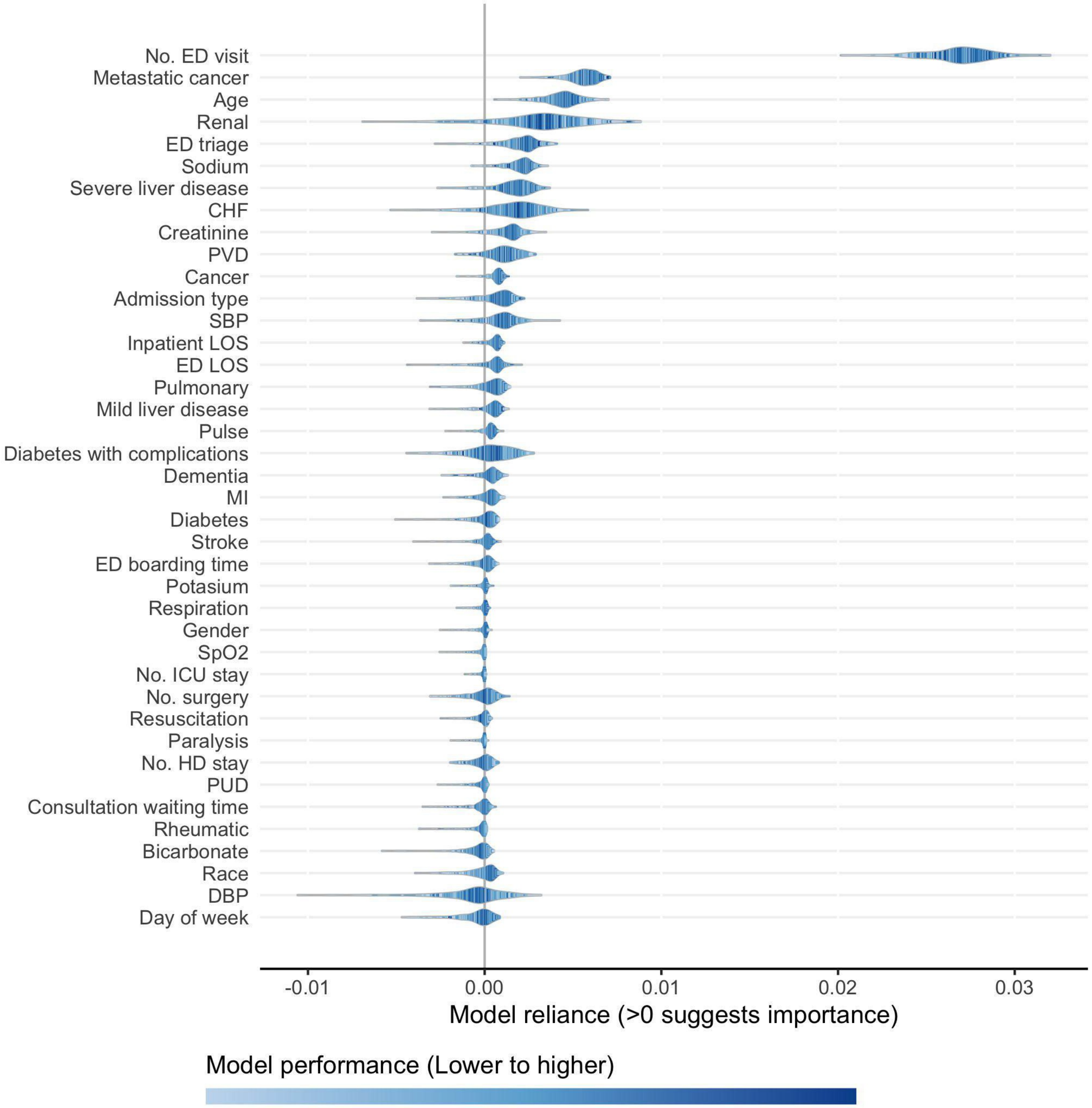

Fig 3. ShapleyVIC values from 350 nearly optimal logistic regression models, arranged in descending order of average ShapleyVIC values. Ventilation had zero ShapleyVIC value in all 350 models analyzed and was therefore not plotted. Number of ED visit is within 6 months before current inpatient stay. Number of surgery, ICU stay and HD stay are within 1 year before current inpatient stay. CHF: Congestive heart failure; DBP: diastolic blood pressure; ED: emergency department; HD: high dependency ward; ICU: intensive care unit; LOS: length of stay; MI: Myocardial infarction; PVD: Peripheral vascular disease; PUD: Peptic ulcer disease; SBP: systolic blood pressure; SpO_2: blood oxygen saturation.

https://doi.org/10.1371/journal.pdig.0000062.g003

resulted in a wider 95% PI and hence a higher uncertainty of its overall importance. Similarly, although admission type had similar overall importance (indicated by the average ShapleyVIC value) as neighboring variables (cancer and systolic blood pressure), the wider spread and long left tail for admission type resulted in a 95% PI containing zero, indicating a non-significant

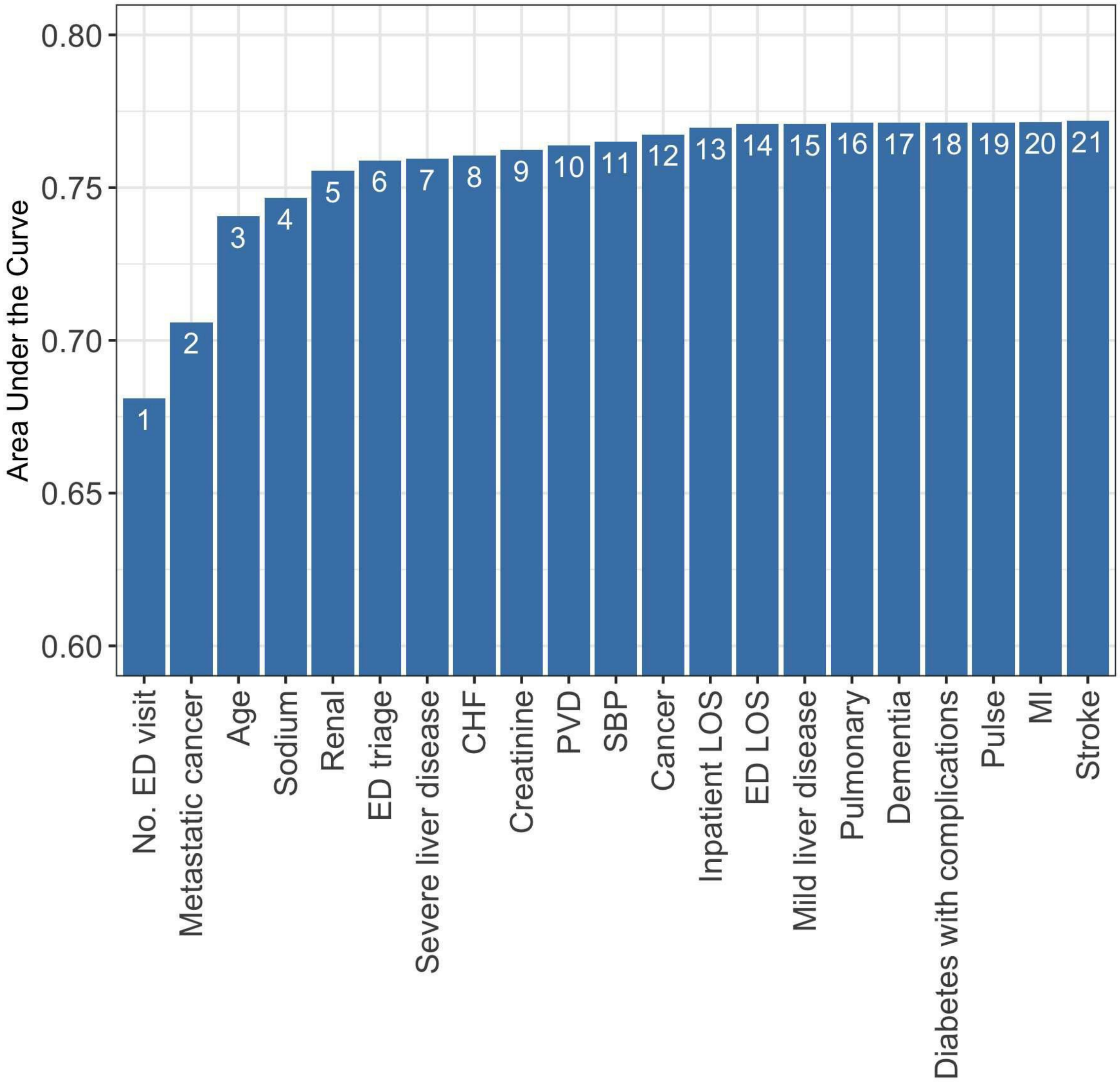

Fig 4. Parsimony plot on the validation set using ShapleyVIC variable ranking, based on 21 variables with significant average ShapleyVIC values. Note that important contributors to model performance, e.g., age and metastatic cancer, are ranked higher by ShapleyVIC than compared to the random-forest-based ranking in Fig 1. Number of ED visit is within 6 months before current inpatient stay. CHF: Congestive heart failure; ED: emergency department; LOS: length of stay; MI: Myocardial infarction; PVD: Peripheral vascular disease; SBP: systolic blood pressure.

https://doi.org/10.1371/journal.pdig.0000062.g004

overall importance for this variable. Among the 41 candidate variables, 20 variables had non-significant overall importance and were excluded from subsequent model building steps.

To rank variables using ShapleyVIC values, we ranked the 41 variables within each of the 350 nearly optimal model, averaged across models to generate an ensemble ranking for all variables and subsequently focused on the 21 variables with significant overall importance. Unlike the parsimony plot from RF-based variable ranking (see Fig 1), the parsimony plot based on the ensemble ranking (see Fig 4) increased smoothly until the 6th variable entered the model, and the increments in model performance became small afterwards. The parsimony plot suggested a feasible model, Model 2, also with 6 variables: number of ED visits in past 6 months, metastatic cancer, age, sodium, renal disease, and ED triage. Four of these 6 variables were in common with the final models built using RF, but now metastatic cancer was selected

Table 4. Scoring table of Model 2 generated using AutoScore, with variables ranked by ShapleyVIC.

Variable	Interval	Point Model 2
Number of ED visits in past 6 months	<1	0
	[1, 3)	14
	> = 3	33
Metastatic cancer	No	0
	Yes	21
Age (years)	<25	0
	[25, 45)	4
	[45, 75)	12
	[75, 85)	15
	> = 85	19
Sodium (mmol/L)	<125	10
	[125, 130)	7
	[130, 135)	4
	> = 135	0
Renal disease	No	0
	Yes	8
ED triage	P1	9
	P2	6
	P3 and P4	0

"[A, B)" indicates an interval inclusive of the lower limit and exclusive of the upper limit.

"—" indicates variables not included in a model.

ED: Emergency department; LOS: length of stay.

https://doi.org/10.1371/journal.pdig.0000062.t004

automatically without the need for manual inspection. A detailed description of steps and corresponding R codes for deriving Model 2 is provided online (see section "AutoScore-ShapleyVIC workflow" of the ShapleyVIC guidebook, available at https://github.com/nliulab/ShapleyVIC).

As a sensitivity analysis, we also generated a parsimony plot using the ensemble ranking of all 41 variables, visualized in Fig A in S1 Text. Although some variables with non-significant overall importance ranked higher than those with significant overall importance (e.g., admission type ranked 14th among all 41 variables, higher than ED LOS and the next seven variables in Fig 4), these non-significant variables had small incremental change to model performance and did not affect the development of Model 2. Hence, the exclusion of variables with non-significant overall importance simplified model building steps without negative impact on the performance of the final model. The scoring table of Model 2 (after fine-tuning) is shown in Table 4, and the AUC evaluated on the test set (0.756, 95% CI: 0.751–0.760) was comparable to that of the 16-variable Model 1B and significantly higher than Model 1A or the LACE index (see Table 2). When evaluated at the optimal threshold identified in the ROC analysis, Model 2 had higher accuracy than Models 1A, 1B or the LACE index (see Table 2).

Discussion

Identifying important variables to the outcome is crucial for developing well-performing and interpretable prediction models [19], especially when developing clinical risk scores based on the simple structure of logistic regression models. The recently developed ShapleyVIC method [18] comprehensively assesses variable importance to accurate predictions, and quantifies the variability in importance to facilitate rigorous statistical inference. In this study, we propose ShapleyVIC as a variable selection tool for developing clinical risk scores, and illustrate its application for binary outcomes in conjunction with the modularized AutoScore framework [13] by developing a risk score for unplanned readmission or death within 30 days after hospital discharge. Based completely on the score-generating logistic regression models, ShapleyVIC avoids bias in importance assessments due to the choice of variable ranking methods, and presents rich statistics on variable contributions through effective visualizations to support in-depth interpretation and guide variable selection. The ShapleyVIC-generated model used 6 of the 41 candidate variables, which outperformed the widely used LACE index and had comparable performance as a 16-variable model developed from RF-based variable ranking. Our work makes a novel contribution to the development of clinical risk scoring systems by providing an effective and robust variable selection method that is supported by statistical assessments of variable importance and is tailored to score-generating regression models.

Models that are ‘good enough’ can be as relevant in real-life clinical applications as the ‘best-performing’ model, especially with respect to practical considerations such as clinical meaningfulness and costs [17,18]. The variable importance cloud proposed by Dong and Rudin [17] was the first to formally extend variable importance to such a group of models with nearly optimal performance. ShapleyVIC creates an ensemble of variable importance measures from a group of models using the state-of-the-art Shapley-based interpretable machine learning method, explicitly evaluates the variability of variable importance across models to support formal assessment of overall importance, and conveys such information through effective visualizations. In this work, we further propose a disciplined approach that takes into account such rich information on variable importance when ranking variables, and demonstrate its easy integration with the existing AutoScore framework [13] for automated development of scoring systems. In addition to using the average ShapleyVIC values across models as a measure of overall variable importance, we propose to use them as a screening tool to exclude candidate variables that are less useful.

Although machine learning methods (e.g., RF, XGBoost and neural networks) can be more efficient than traditional variable selection methods [12,13,20], they often assume complex non-linear and/or tree-based structures that differ drastically from the linear structure assumed by score-generating logistic regression. Our application highlights such misalignment, where an important contributor (metastatic cancer) to the scoring model was only ranked 16th among all 41 variables by the RF, and the 4th-ranking variable (ED boarding time) by the RF had minimal incremental contribution to the performance of the scoring model. Such issues are not limited to the RF but are generally shared by machine learning methods. For example, similar issues were observed for XGBoost-based variable ranking in an additional study (see Fig B in S1 Text for the parsimony plot). While it may be relevant to explore other machine learning methods (e.g., neural networks) for variable ranking, it can be challenging to train such models (e.g., to determine the number of layers and number of nodes per layer in a neural network) to reasonably evaluate variable importance, making them practically less useful. Hence, we did not include additional variable ranking methods in our comparative evaluations.

The misalignment between the RF-based variable ranking and variable contributions to the scoring model is less problematic to the AutoScore framework, as it can be observed from the parsimony plot and handled by manual adjustment to the variable list, e.g., by excluding ED boarding time from Model 1A and adding metastatic cancer to build a new 6-variable model with comparable performance (AUC = 0.756, 95% CI: 0.751–0.760) to the 16-variable Model 1B. ShapleyVIC avoids such subjective assessments by providing an objective and data-driven ranking approach that is tailored to the score-generating logistic regression. Using an ensemble variable ranking across 350 well-performing logistic regression models, ShapleyVIC successfully assigned high ranks to all important contributors (including metastatic cancer), and excluded ED boarding time with other 19 variables that had non-significant overall importance. When the final list of variables is selected and the scoring table is derived using the AutoScore workflow, predicting the outcome for a new patient is simple: clinicians can first compute a total score for the patient by summating the points corresponding to each variable value, and then predict the outcome if the total score exceeds a threshold (e.g., a data-driven threshold based on ROC analysis, or an adjusted threshold corresponding to a satisfactory sensitivity or specificity).

Model 2 developed using ShapleyVIC-based variable ranking used 6 variables, among which 5 variables (i.e., ED triage, age, number of ED visits in past 6 months, metastatic cancer and renal disease) are easily collected at ED registration or retrieved from hospital database. The 6th variable, sodium level at ED, is less convenient to collect but may be imputed or assigned 0 point in the score if not available. Hence, Model 2 is feasible for clinical application and so is the 6-variable Model 1A for similar reasons, but the 16-variable Model 1B can be difficult to implement in clinical applications. Although the widely used LACE index has been recognized as a 4-variable scoring model, it is worth noting that one of the predictors, the Charlson comorbidity index (CCI), aggregates information across 17 medical conditions, therefore may not be convenient for quick risk stratification. Despite the simplicity of Model 2, it outperformed Model 1A and the LACE index, and had comparable performance as Model 1B. These demonstrate the ability of the integrated AutoScore-ShapleyVIC framework in developing well-performing and sparse risk scoring models.

Compared with the LACE index, Model 2 only requires information on 2 of the 17 comorbidities (i.e., metastatic cancer and renal disease), uses ED triage instead of inpatient LOS, and requires additional information on patient age and sodium level at ED. We find the use of ED triage but not inpatient LOS in Model 2 reasonable in our setting, because the full cohort consists of acute admissions (i.e., inpatient admissions after ED visits), and ED triage is an important predictor for short-term clinical outcomes such as 30-day mortality or readmission [21]. The inclusion of age in Model 2 is not surprising, as old age generally associates with alleviated risk of adverse clinical outcomes, but the clinical basis of the use of sodium level may warrant further investigation. The focus on metastatic cancer and renal disease among the 17 comorbidities is consistent with a recent study of time to emergency readmission using a similar cohort to that in our study [22], which developed a 6-variable risk score using number of ED visits in previous year, age, cancer history, renal disease history and inpatient measures of creatinine and album. Three variables in Model 1B (i.e., number of ED visits in past 6 months, inpatient LOS and metastatic cancer) were also included in the LACE index. Model 1B did not include renal disease but included creatinine level at ED, which is reflective of renal function to some extent [23]. In addition to patient age, Model 1B also requires additional information on ED admission and several laboratory tests and vital signs at ED (including sodium used in Model 2). However, since the contribution of some of the variables in Model 1B to predictive performance has been questionable (e.g., see the 4th and 7-15th variables in Fig 1), we refrain from further discussing the clinical implications for these variables. Readers should also note that scoring models discussed above were developed for illustrative purpose and should not be use in clinical applications without further investigations and validation. Moreover, when developing risk prediction models for composite outcomes, e.g., in our study a composite of readmission or death, it is useful to separately evaluate the model for each event for improved interpretation of the prediction model and a more comprehensive evaluation of predictive performance.

Interestingly, additional analyses found that when using the same variables, RF had a slightly lower AUC than Model 1B (AUC = 0.754, 95% CI: 0.749–0.758) and a much lower AUC than Model 2 (AUC = 0.663, 95% CI: 0.659–0.668). This again highlights the drastic difference between RF and logistic regression models and their perception on variable importance, and that complex and robust machine learning models do not necessarily outperform simple scoring models. However, compared to variable ranking based on a single RF, ShapleyVIC requires much longer run-time and larger memory space, which would benefit from the use of high-performance computers and parallel computing (option available in the ShapleyVIC R package [24]). Another limitation of ShapleyVIC is that the sampled set of models generated using our pragmatic sampling approach may not fully represent the entire set of near-optimal models (formally referred to as the Rashomon set) [18], which remains a challenge in this area of research [1,17]. In addition, although collinearity was not present in our example (where generalized variance inflation factor [VIF] for all 41 candidate variables were below 2), it is generally an unresolved difficulty in variable importance assessments [17,18,25–27]. Our pragmatic solution of using absolute SAGE values as model reliance measures for variables with VIF>2 worked well in a previous empirical experiment [18], and future work aims to devise more formal solutions. In current study, we adopted the holdout method that randomly splits the full cohort into three datasets for model development and validation, which is a common practice when working with large datasets. Future work will develop and validate an alternative workflow using cross-validation to accommodate smaller datasets.

Our work contributes to the recent emphasis on interpretability and transparency of prediction models for high-stakes decision making [28], by devising a robust and interpretable approach for developing clinical scores. Unlike existing statistical or machine learning methods for variable selection, which focus on ranking variables by their importance to the prediction, ShapleyVIC rigorously evaluates variable contributions to the scoring model and visualizes the findings to enable a robust and fully transparent variable ranking procedure. The ShapleyVIC-based variable ranking, which is tailored to the score-generating logistic regression, is easily applied to the AutoScore framework for generating sparse scoring models with the flexibility to adjust for expert opinions and practical conventions. ShapleyVIC and AutoScore combines into an integrated approach for future interpretable machine learning research in clinical applications, which provides a disciplined solution to detailed assessment of variable importance and transparent development of clinical risk scores, and is easily implemented by executing a few commands from the respective R packages [24,29]. Although we have illustrated our approach for predicting early death or unplanned readmission, it is not limited to any specific clinical application. For example, AutoScore has recently been used to derive a risk score for ED triage that outperformed several existing indices [14], and to provide a well-performing risk score for survival after return of spontaneous circulation in out-of-hospital cardiac arrest that is easier to implement than existing alternatives [30]. Future work will explore the performance of AutoScore-ShapleyVIC in such diverse application scenarios, and validate its performance in risk score derivation against other score generating systems and existing prediction models. Current work describes the implementation of our proposed scoring system for binary outcomes, but the model-agnostic property of ShapleyVIC [18] and the recent extension of the AutoScore framework to survival outcomes [31] enables the use of our approach for a wider range of applications.

Methods

AutoScore framework for developing risk scoring models

The AutoScore framework provides an automated procedure for developing risk scoring models for binary outcomes. Interested readers may refer to the original paper [13] and the R package documentation [29] for detailed description of AutoScore methodology, and to a clinical application for an example use case [14]. In brief, AutoScore divides data into training, validation and test sets (typically consisting of 70%, 10% and 20% of total sample size, respectively), and ranks all candidate variables based on their importance to a RF trained on the training set. Continuous variables are then automatically categorized (using percentiles or k-means clustering) [13] for preferable clinical interpretation and to account for potential non-linear relationships between predictors and outcomes. AutoScore creates scoring models based on the logistic regression, where coefficients are scaled and rounded to non-negative integer values for convenient calculation. Following the RF-based variable ranking, AutoScore grows the scoring model by adding one variable each time and inspects the resulting improvement in model performance (measured by the AUC on the validation set) using a parsimony plot. The final model is often determined by selecting the top few variables where a reasonable AUC is reached and inclusion of an additional variable only results in a small increment (e.g., <1%) to AUC. After selecting the final list of variables, cut-off values used to categorize continuous variables may be fine-tuned to suit clinical practice and conventions, and the finalized model is evaluated on the test set.

Shapley variable importance cloud (ShapleyVIC)

In practical prediction tasks, researchers are often willing to select a simpler and interpretable model even when its performance is marginally lower than a more complex model. Hence, conventional variable importance assessments based on a single best model can be limiting. ShapleyVIC [18] is suitable for such purposes and assesses the overall contribution of variables by studying a group of models with nearly optimal performance. In the case of developing scoring models for binary outcomes, nearly optimal models can be characterized by vectors of regression coefficients corresponding to logistic loss that exceeds the minimum logistic loss by no more than a selected threshold, for which 5% is a reasonable choice but may be adjusted based on practical needs and considerations.

ShapleyVIC consists of three general steps. First, a suitable number (e.g., 350) of nearly optimal models (defined using the selected threshold for model loss) are sampled from a multivariable normal distribution centered at the regression coefficients of the optimal logistic regression model. Specifically, due to the difficulty in mathematically defining the boundary of this multivariable normal distribution corresponding to the desirable range of model loss, we proposed a rejection sampling approach to empirically explore the distribution, with two tunable parameters to control the range explored. The sampling steps and instructions on parameter tuning are described in detail in the original paper [18] and implemented in the ShapleyVIC package [24]. Next, ShapleyVIC measures reliance of each nearly optimal model on variables using the Shapley additive global importance (SAGE) method [26], which measures variable impact on a model using Shapley values and is closely related to the commonly used SHAP method. As explained in the original paper on ShapleyVIC, we use the absolute SAGE values instead of the unadjusted values to measure model reliance for variables involved in collinearity, identified by VIF>2 from the optimal logistic regression model. Finally, the ShapleyVIC values are pooled across all nearly optimal models using a random-effect meta-analysis approach, which quantifies the overall importance of variables and explicitly evaluates the variability of these measures across well-performing models (in terms of a 95% PI of variable importance for a new model) to support formal statistical inference. The average ShapleyVIC values and 95% PIs are visualized using a bar plot, and the variability of ShapleyVIC values across models is highlighted using a colored violin plot for additional insights.

Ensemble variable ranking from ShapleyVIC

To apply ShapleyVIC in practical development of scoring models based on logistic regression, we use an ensemble variable ranking approach from ShapleyVIC that is tailored to score-generating models and filter out variables that are not likely to make significant contributions. First, we describe our proposed ensemble variable ranking for *d* variables for a single nearly optimal model *f*. Let $\hat{mr}_j^s(f)$ denote the ShapleyVIC value of the *j*-th variable ($j = 1,\ldots,d$) for model *f*, which, as described above, is either the unadjusted SAGE value or the absolute SAGE value depending on the VIF of the variable. The variability of $\hat{mr}_j^s(f)$ can therefore be measured using the standard error of the SAGE value, denoted by $\sigma_j(f)$, that is readily available from the SAGE algorithm. Assuming independent normal distributions for SAGE values [18,26] of two variables (e.g., X_j and X_k) to model *f*, the difference between their ShapleyVIC values is also normally distributed: $\hat{mr}_j^s(f) - \hat{mr}_k^s(f) \sim N(mr_j^s(f) - mr_k^s(f), \sigma_j^2(f) + \sigma_k^2(f))$. We compare all possible pairs of variables, and subsequently rank the *d* variables to model *f* based on the number of times each variable has significantly larger ShapleyVIC value than the other *d*–1 variables. Assigning rank 1 to the most important variable and using the same smallest integer value available for tied ranks, we rank the *d* variables for each model, and use the average rank of each variable across all nearly optimal models to generate an ensemble ranking.

Instead of considering all *d* candidate variables in subsequent model building steps, we propose to filter out variables that are not likely important contributors to accurate predictions based on their overall importance, corresponding to variables where the 95% PI for the average ShapleyVIC value contains or is entirely below zero. In this work, we applied the proposed ensemble variable ranking to the AutoScore framework, using it to replace RF when ranking variables.

Study design and data preparation

Our empirical experiment was based on a retrospective cohort study of patients who visited the ED of Singapore General Hospital in years 2009 to 2017 and were subsequently admitted to the hospital. The outcome of interest was readmission or death within 30 days after discharge from hospital. Since data was collected from ED, all readmissions in the cohort were unplanned readmissions. Information on patient demographics, ED administration, inpatient admission, healthcare utilization, clinical tests, vital signs, and comorbidities was extracted from the hospital electronic health record system. A full list of variables is provided in Table 1. ED triage was evaluated using the Patient Acuity Category Scale [21], the current triage system used in Singapore. Admission type refers to the four types of wards in the hospital with different costs [32], which partially reflects the socio-economic status of patients. We excluded patients who were not Singapore residents, died in hospital, aged below 21 or had missing information on ED boarding time or consultation waiting time. The final cohort was split into training (70%), validation (10%) and testing sets (20%), and missing values for vital signs or clinical tests were imputed using the median value in the training set. To compute the LACE index, we computed the CCI as described in a previous work [33].

Ethics

This study was approved by Singapore Health Services' Centralized Institutional Review Board (CIRB 2021/2122), and a waiver of consent was granted for electronic health record data collection.

Supporting information

S1 Text. Supplementary table and figures.
(PDF)

Author Contributions

Conceptualization: Yilin Ning, Nan Liu.

Data curation: Yilin Ning.

Formal analysis: Yilin Ning, Siqi Li.

Investigation: Yilin Ning, Marcus Eng Hock Ong, Feng Xie, Bibhas Chakraborty, Daniel Shu Wei Ting, Nan Liu.

Methodology: Yilin Ning, Siqi Li, Marcus Eng Hock Ong, Feng Xie, Bibhas Chakraborty, Daniel Shu Wei Ting, Nan Liu.

Project administration: Yilin Ning, Nan Liu.

Software: Yilin Ning, Siqi Li.

Supervision: Nan Liu.

Validation: Yilin Ning, Siqi Li, Marcus Eng Hock Ong, Feng Xie, Bibhas Chakraborty, Daniel Shu Wei Ting, Nan Liu.

Writing – original draft: Yilin Ning.

Writing – review & editing: Yilin Ning, Siqi Li, Marcus Eng Hock Ong, Feng Xie, Bibhas Chakraborty, Daniel Shu Wei Ting, Nan Liu.

References

1. Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. Stat Surv. 2022; 16: 1–85. https://doi.org/10.1214/21-SS133
2. van Walraven C, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. Can Med Assoc J. 2010; 182: 551 LP– 557. https://doi.org/10.1503/cmaj.091117 PMID: 20194559
3. Robinson R, Hudali T. The HOSPITAL score and LACE index as predictors of 30 day readmission in a retrospective study at a university-affiliated community hospital. PeerJ. 2017; 5: e3137–e3137. https://doi.org/10.7717/peerj.3137 PMID: 28367375
4. Damery S, Combes G. Evaluating the predictive strength of the LACE index in identifying patients at high risk of hospital readmission following an inpatient episode: a retrospective cohort study. BMJ Open. 2017; 7: e016921. https://doi.org/10.1136/bmjopen-2017-016921 PMID: 28710226
5. Su M-C, Chen Y-C, Huang M-S, Lin Y-H, Lin L-H, Chang H-T, et al. LACE Score-Based Risk Management Tool for Long-Term Home Care Patients: A Proof-of-Concept Study in Taiwan. Int J Environ Res Public Health. 2021; 18. https://doi.org/10.3390/ijerph18031135 PMID: 33525331
6. Tong L, Erdmann C, Daldalian M, Li J, Esposito T. Comparison of predictive modeling approaches for 30-day all-cause non-elective readmission risk. BMC Med Res Methodol. 2016; 16: 26. https://doi.org/10.1186/s12874-016-0128-0 PMID: 26920363
7. Artetxe A, Beristain A, Graña M. Predictive models for hospital readmission risk: A systematic review of methods. Comput Methods Programs Biomed. 2018; 164: 49–64. https://doi.org/10.1016/j.cmpb.2018.06.006 PMID: 30195431
8. Wang S, Zhu X. Predictive Modeling of Hospital Readmission: Challenges and Solutions. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2021. p. 1. https://doi.org/10.1109/TCBB.2021.3089682 PMID: 34133285
9. Zhang Z, Zhang H, Khanal MK. Development of scoring system for risk stratification in clinical medicine: a step-by-step tutorial. Annals of translational medicine. 2017. p. 436. https://doi.org/10.21037/atm.2017.08.22 PMID: 29201888
10. Smith G. Step away from stepwise. J Big Data. 2018; 5: 32. https://doi.org/10.1186/s40537-018-0143-6
11. Su W, Bogdan M, Candes E. False discoveries occur early on the Lasso path. Ann Stat. 2017; 45: 2133–2150.

12. Shadmi E, Flaks-Manov N, Hoshen M, Goldman O, Bitterman H, Balicer RD. Predicting 30-Day Readmissions With Preadmission Electronic Health Record Data. Med Care. 2015;53. https://doi.org/10.1097/MLR.0000000000000315 PMID: 25634089

13. 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 Informatics. 2020; 8: e21798. https://doi.org/10.2196/21798 PMID: 33084589

14. 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: e2118467–e2118467. https://doi.org/10.1001/jamanetworkopen.2021.18467 PMID: 34448870

15. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy. 2021; 23: 18. https://doi.org/10.3390/E23010018 PMID: 33375658

16. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. pp. 4768–4777.

17. Dong J, Rudin C. Exploring the cloud of variable importance for the set of all good models. Nat Mach Intell. 2020; 2: 810–824. https://doi.org/10.1038/s42256-020-00264-0

18. Ning Y, Ong MEH, Chakraborty B, Goldstein BA, Ting DSW, Vaughan R, et al. Shapley variable importance clouds for interpretable machine learning. Patterns. 2022; 100452. https://doi.org/10.1016/j.patter.2022.100452 PMID: 35465224

19. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. J Mach Learn Res. 2003; 3: 1157–1182. https://doi.org/10.5555/944919.944968

20. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform. 2008; 77: 81–97. https://doi.org/10.1016/j.ijmedinf.2006.11.006 PMID: 17188928

21. Fong RY, Glen WSS, Mohamed Jamil AK, Tam WWS, Kowitlawakul Y. Comparison of the Emergency Severity Index versus the Patient Acuity Category Scale in an emergency setting. Int Emerg Nurs. 2018; 41: 13–18. https://doi.org/10.1016/j.ienj.2018.05.001 PMID: 29887281

22. Xie F, Liu N, Yan L, Ning Y, Lim KK, Gong C, et al. Development and validation of an interpretable machine learning scoring tool for estimating time to emergency readmissions. eClinicalMedicine. 2022; 45: 101315. https://doi.org/10.1016/j.eclinm.2022.101315 PMID: 35284804

23. Levey A, Perrone R, Madias N. Serum creatinine and renal function. Annu Rev Med. 1988; 39: 465–490. https://doi.org/10.1146/annurev.me.39.020188.002341 PMID: 3285786

24. Ning Y, Li S. ShapleyVIC: Shapley Variable Importance Cloud for Interpretable Machine Learning. 2022. Available: https://github.com/nliulab/ShapleyVIC

25. Fisher A, Rudin C, Dominici F. 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. 2019; 20: 1–81. PMID: 34335110

26. Covert IC, Lundberg S, Lee S-I. Understanding Global Feature Contributions With Additive Importance Measures. Advances in Neural Information Processing Systems 33 (NeurIPS 2020). 2020. pp. 17212–17223.

27. Covert I, Lee S-I. Improving KernelSHAP: Practical Shapley Value Estimation Using Linear Regression. Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. PMLR; 2021. pp. 3457–3465.

28. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019; 1: 206–215. https://doi.org/10.1038/s42256-019-0048-x PMID: 35603010

29. Xie F, Ning Y, Yuan H, Saffari E, Chakraborty B, Liu N. AutoScore: An Interpretable Machine Learning-Based Automatic Clinical Score Generator. 2021. Available: https://cran.r-project.org/package=AutoScore

30. 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: 126–133. https://doi.org/10.1016/j.resuscitation.2021.11.029 PMID: 34843878

31. Xie F, Ning Y, Yuan H, Goldstein BA, Ong MEH, Liu N, et al. AutoScore-Survival: Developing interpretable machine learning-based time-to-event scores with right-censored survival data. J Biomed Inform. 2021; 125: 103959. https://doi.org/10.1016/j.jbi.2021.103959 PMID: 34826628

32. Singapore General Hospital. Type of Wards. Available: https://www.sgh.com.sg/patient-care/inpatient-day-surgery/type-of-wards-singapore-general-hospital

33. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A New Method of Classifying Prognostic in Longitudinal Studies: Development and Validation. J Chronic Dis. 1987; 40: 373–383. https://doi.org/10.1016/0021-9681(87)90171-8 PMID: 3558716