Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study

Romain Pirracchio, Maya L Petersen, Marco Carone, Matthieu Resche Rigon, Sylvie Chevret, Mark J van der Laan

Lancet Respir Med 2015;
3: 42–52

Published Online
November 24, 2014
http://dx.doi.org/10.1016/
S2213-2600(14)70239-5

See Comment page 5

Division of Biostatistics, School of Public Health, University of California, Berkeley, CA, USA (R Pirracchio MD, M L Petersen MD, Prof M J van der Laan PhD); **Service de Biostatistique et Information Médicale, Unité INSERM 1153, Equipe ECSTRA, Hôpital Saint Louis, Paris, France** (R Pirracchio, M R Rigon MD, Prof S Chevret MD)**; Service d'Anesthésie-Réanimation, Hôpital Européen Georges Pompidou, Paris, France** (R Pirracchio)**; and Department of Biostatistics, School of Public Health, University of Washington, Seattle, WA, USA** (M Carone PhD)

Correspondence to:
Dr Romain Pirracchio, Service d'Anesthésie-Réanimation, Hôpital Européen Georges Pompidou, Paris 75015, France
romain.pirracchio@egp.aphp.fr

Summary

Background Improved mortality prediction for patients in intensive care units is a big challenge. Many severity scores have been proposed, but findings of validation studies have shown that they are not adequately calibrated. The Super ICU Learner Algorithm (SICULA), an ensemble machine learning technique that uses multiple learning algorithms to obtain better prediction performance, does at least as well as the best member of its library. We aimed to assess whether the Super Learner could provide a new mortality prediction algorithm for patients in intensive care units, and to assess its performance compared with other scoring systems.

Methods From January, 2001, to December, 2008, we used the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database (version 26) including all patients admitted to an intensive care unit at the Beth Israel Deaconess Medical Centre, Boston, MA, USA. We assessed the calibration, discrimination, and risk classification of predicted hospital mortality based on Super Learner compared with SAPS-II, APACHE-II, and SOFA. We calculated performance measures with cross-validation to avoid making biased assessments. Our proposed score was then externally validated on a dataset of 200 randomly selected patients admitted at the intensive care unit of Hôpital Européen Georges-Pompidou, Paris, France, between Sept 1, 2013, and June, 30, 2014. The primary outcome was hospital mortality. The explanatory variables were the same as those included in the SAPS II score.

Findings 24 508 patients were included, with median SAPS-II of 38 (IQR 27–51) and median SOFA of 5 (IQR 2–8). 3002 of 24 508 (12%) patients died in the Beth Israel Deaconess Medical Centre. We produced two sets of predictions based on the Super Learner; the first based on the 17 variables as they appear in the SAPS-II score (SL1), and the second, on the original, untransformed variables (SL2). The two versions yielded average predicted probabilities of death of 0·12 (IQR 0·02–0·16) and 0·13 (0·01–0·19), whereas the corresponding value for SOFA was 0·12 (0·05–0·15) and for SAPS-II 0·30 (0·08–0·48). The cross-validated area under the receiver operating characteristic curve (AUROC) for SAPS-II was 0·78 (95% CI 0·77–0·78) and 0·71 (0·70–0·72) for SOFA. Super Learner had an AUROC of 0·85 (0·84–0·85) when the explanatory variables were categorised as in SAPS-II, and of 0·88 (0·87–0·89) when the same explanatory variables were included without any transformation. Additionally, Super Learner showed better calibration properties than previous score systems. On the external validation dataset, the AUROC was 0·94 (0·90–0·98) and calibration properties were good.

Interpretation Compared with conventional severity scores, Super Learner offers improved performance for predicting hospital mortality in patients in intensive care units. A user-friendly implementation is available online and should be useful for clinicians seeking to validate our score.

Funding Fulbright Foundation, Assistance Publique–Hôpitaux de Paris, Doris Duke Clinical Scientist Development Award, and the NIH.

Introduction

The burden of care for critically ill patients is huge. In the USA, the cost of care for critically ill patients accounts for nearly 1% of the gross domestic product, and although less than 10% of hospital beds are found in intensive care units (ICU), ICU departments account for 22% of total hospital costs.[1] In the UK, the cost of intensive care is estimated to be £541 million per year, which represents 0·6% of National Health Service expenditures.[2] During 2009–12, the average hospital mortality rate for patients in ICU was estimated to be 11–12%.[3] Prediction of mortality in patients in ICU is crucial for the assessment of severity of illness and adjudication of the value of novel treatments, interventions, and health-care policies. In the past 30 years, a big effort has been made in modelling the risk of death in patients in ICU. Several severity scores have been developed with the objective of predicting hospital mortality from baseline patient characteristics.

The first scores proposed with the Acute Physiology and Chronic Health Evaluation (APACHE),[4] APACHE- II,[5] and Simplified Acute Physiology Score (SAPS),[6] relied on subjective methods for variable selection, namely relying on a panel of experts to select and assign weights to variables according to perceived relevance for mortality prediction. Further scores, such as the SAPS-II,[7] were subsequently developed with statistical modelling techniques.[7–10] Up to now, the SAPS-II[7] and APACHE-II[5] scores remain the most widely used in clinical practice. However, since first being published, they have been modified several times to improve their predictive performance.[9,10] These scores discriminate survivors and non-survivors well. However, data from several external validation studies done in various countries have suggested that the most recent versions of SAPS and APACHE are not adequately calibrated, which means that they fail to accurately predict the actual probability of death.[11,12] Locally customised variants of these scores have also been developed to incorporate regional variations. For instance, versions of the SAPS score have been specifically tailored to France, southern Europe, and Mediterranean countries, and to central and western Europe.[10,13,14] Despite these extensions of SAPS, predicted hospital mortality remains generally overestimated.[11,12,15–17]

Most ICU severity scores rely on a logistic regression model. Such models impose stringent constraints on the association between explanatory variables and risk of death. For instance, main-term logistic regression typically relies on a linear and additive relationship between a pre-specified transformation of the mean outcome and its predictors. In view of the complex processes underlying death in patients in ICU, such an assumption might be unrealistic, and predictive power might be low if an incorrect parametric model is used as opposed to a more flexible option. On the contrary, if the assumed parametric model is correct, it will generally provide the best prediction, at least in large samples. Hence, the poor calibration of present severity scores might be, to a large extent, a consequence of the misspecification of the underlying statistical model rather than to the choice of variables included in this model. We aimed to assess whether a more flexible statistical approach, namely the Super Learner, could improve ICU mortality prediction compared with conventional methods without needing to include additional variables in the scoring procedure.

Methods

Study design and participants

The MIMIC-II study[18–20] includes all patients admitted to an ICU at the Beth Israel Deaconess Medical Centre (BIDMC), Boston, MA, USA, since 2001. Patient recruitment is still in progress. In this study, we only included data from MIMIC-II version 26 (2001–08) for adult patients (aged >15 years) in ICU.

The BIDMC is a 620-bed tertiary academic medical centre and a level one trauma centre with 77 critical care beds. The ICUs at the BIDMC are closed (ie, the intensivists are responsible for patient care, not the physician referring the patient to the ICU), with continuous in-house supervision of patient care by an intensivist. These ICUs include medical, trauma-surgical, coronary, cardiac surgery recovery, and medicosurgical critical care units.

All consecutive patients were included in the MIMIC-II database. Staff were not involved with data acquisition and did not interfere with the clinical care of patients or methods of monitoring. We included only patients with one ICU admission per hospital stay. We collected two categories of data: clinical data, aggregated from ICU information systems and hospital archives, and high-resolution physiological data (waveforms and time series of derived physiological measurements), recorded on bedside monitors. Clinical data were obtained from the CareVue clinical information system (models M2331A and M1215A, Philips Healthcare, Andover, MA, USA) deployed in all study ICUs, and from hospital electronic archives. The data included time-stamped nurse-verified physiological measurements (eg, measurements of heart rate, arterial blood pressure, and pulmonary artery pressure every hour), nurses' and respiratory therapists' progress notes, continuous intravenous drip drugs, fluid balances, patient demographics, interpretations of imaging studies, physician orders, discharge summaries, and International Classification of Diseases-9 (ICD-9) codes. Comprehensive diagnostic laboratory results (eg, blood chemistry, complete blood counts, arterial blood gases, and microbiology results) were obtained from the patient's entire hospital stay including periods outside the ICU. In the present study, we focused exclusively on outcome variables (specifically ICU and hospital mortality) and variables included in the SAPS-II[7] and SOFA scores.[21]

This study was approved by the institutional review boards of BIDMC and the Massachusetts Institute of Technology (Cambridge, MA, USA). Requirement for individual patient consent was waived because the study did not affect clinical care and all protected health information was de-identified. De-identification was done in compliance with Health Insurance Portability and Accountability Act (HIPAA) standards to facilitate public access to MIMIC-II. Deletion of protected health information from structured data sources (eg, database fields that provide patient name or date of birth) was direct and systematic. Additionally, protected health information was removed from the discharge summaries and diagnostic reports and the roughly 700 000 free-text nursing and respiratory notes in MIMIC-II with an automated algorithm previously shown to outperform clinicians in detecting protected health information.[22]

Outcomes and procedures

The primary outcome measure was hospital mortality. The data recorded within the first 24 h after ICU admission were extracted separately from the MIMIC-II (version 26) database and used to compute two of the most widely used severity scores, namely the SAPS-II[7] and SOFA[21] scores. Individual mortality prediction for the SAPS-II score was calculated as defined by its authors:[7]

$$\log\left[\frac{\text{pr(death)}}{1-\text{pr(death)}}\right] = \begin{array}{l} -7\cdot7631+0\cdot0737\times\text{SAPS-II}+ \\ 0\cdot9971\times\log(1+\text{SAPS-II}) \end{array}$$

Additionally, we developed a new version of the SAPS-II score, by fitting a main-term logistic regression model to our data by use of the same explanatory variables as those used in the original SAPS-II score.[7] The same procedure was used to build a new version of the APACHE-II score.[5] Finally, we computed SOFA score for all participants because it is sometimes used in clinical practice as a proxy for outcome prediction.[21] We obtained mortality prediction based on the SOFA score by regressing hospital mortality on the SOFA score with a main-term logistic regression. We compared these two algorithms for mortality prediction with our Super Learner-based proposal.

The Super Learner has been proposed as a method for selecting via cross-validation the optimum regression algorithm among all weighted combinations of a set of candidate algorithms (ie, the library; appendix pp 3–4).[23–25] To implement the Super Learner, a user needs to provide a customised collection of various data-fitting algorithms and also specify a performance measure (in this study the squared difference between observed and predicted outcomes). The Super Learner then uses V-fold cross-validation to estimate the mean squared prediction error of each algorithm on data not used when building the prediction model, and then selects the convex combination of algorithms that provides the smallest squared prediction error on independent data.

See Online for appendix

	Overall population (n=24 508)	Dead at hospital discharge (n=3002)	Alive at hospital discharge (n=21 506)
Age (years)	65 (51–77)	74 (59–83)	64 (50–76)
Sex (% women)	13 838 (57%)	1607 (54%)	12 231 (57%)
First SAPS	13 (10–17)	18 (14–22)	13 (9–17)
First SAPS-II	38 (27–51)	53 (43–64)	36 (27–49)
First SOFA	5 (2–8)	8 (5–12)	5 (2–8)
Type of admission			
Medical	2453 (10%)	240 (8%)	2213 (10%)
Trauma	7703 (31%)	1055 (35%)	6648 (31%)
Emergency surgery	10 803 (44%)	1583 (53%)	9220 (43%)
Scheduled surgery	3549 (15%)	124 (4%)	3425 (16%)
Type of ICU			
Medical	7488 (31%)	1265 (42%)	6223 (29%)
Medicosurgical	2686 (11%)	347 (12%)	2339 (11%)
Coronary	5285 (22%)	633 (21%)	4652 (22%)
Cardiac surgery recovery	8100 (33%)	664 (22%)	7436 (35%)
Trauma surgical	949 (4%)	93 (3%)	856 (4%)
Heart rate (bpm)	87 (75–100)	92 (78–109)	86 (75–99)
Mean arterial pressure (mm Hg)	81 (70–94)	78 (65–94)	82 (71–94)
Respiratory rate (cpm)	14 (12–20)	18 (14–23)	14 (12–18)
Serum sodium (mmol/L)	139 (136–141)	138 (135–141)	139 (136–141)
Serum potassium (mmol/L)	4·2 (3·8–4·6)	4·2 (3·8–4·8)	4·2 (3·8–4·6)
Serum bicarbonates (mmol/L)	26 (22–28)	24 (20–28)	26 (23–28)
White blood cell count (10^3/mm^3)	10·3 (7·5–14·4)	11·6 (7·9–16·9)	10·2 (7·4–14·1)
$Pa0_2/Fi0_2$	281 (130–447)	174 (90–352)	312 (145–461)
Haematocrit (%)	34·7 (30·4–39)	33·8 (29·8–38)	34·8 (30·5–39·1)
Urea nitrogen (mmol/l)	20 (14–31)	28 (18–46)	19 (13–29)
Bilirubin (μmol/L)	0·6 (0·4–1)	0·7 (0·4–1·5)	0·6 (0·4–0·9)
Hospital length of stay (days)	8 (4–14)	9 (4–17)	8 (4–14)
ICU death (%)	1978 (8%)	1978 (66%)	..

Data are median (IQR) or count (%). SAPS=Simplified Acute Physiology Score. SOFA=Sepsis-related Organ Failure Assessment. ICU=intensive care unit. bpm=beats per minute. cpm=counts per minute.

Table 1: **Baseline characteristics and outcome measures**

Comparison of the 12 algorithms relied on ten-fold cross-validation. We split data into ten mutually exclusive and exhaustive blocks of roughly equal size (appendix p 2). Each algorithm was fitted on nine blocks (the training set) and used to predict mortality for patients in the remaining block (the validation set). We then calculated the mean squared error between predicted and recorded outcomes. This procedure was repeated ten times, with a different block used as validation set every time. Therefore, each finding served exactly once in the validation set and was included in the training set for all other rounds. We aggregated performance measures over all ten iterations, yielding a cross-validated estimate of the mean-squared error (CV-MSE) for each algorithm. A crucial aspect of this approach is that for each iteration, no patient appears in both the training and validation sets. The potential for overfitting, wherein the fit of an algorithm is overly tailored to the available data at the expense of performance on future data, is thereby mitigated because overfitting is more likely when training and validation sets intersect. Candidate algorithms were ranked according to their CV-MSE and the algorithm with least CV-MSE was identified. We then refitted the algorithm with all available data, leading to a prediction rule referred to as the Discrete Super Learner. Subsequently, we computed the prediction rule consisting of the CV-MSE-minimising weighted convex combination of all candidate algorithms and refitted on all data (ie, the Super Learner combination algorithm).[25] Finally, we assessed the performance of the Super Learner combination algorithm with an additional layer of cross validation; the entire procedure was run in turn on each 9/10th of the data, and performance measures described below were assessed on the remaining validation set and averaged across the ten validation sets.

Theoretical data suggest that, to optimise the performance of the resulting algorithm, the inputted library should include as many algorithms as possible. In this study, the library size was limited to 12 algorithms (appendix pp 3–4) for computational reasons. Of these 12 algorithms, some were parametric, such as logistic regression or related methods classically used for ICU scoring systems, and some were non-parametric—ie, they imposed only minimum constraints on the underlying data distribution. In the present study, we chose the library to include most of the parametric (including regression models with various combinations of main and interaction terms as well as splines, and fitted using maximum likelihood with or without penalisation) and non-parametric algorithms previously assessed for the prediction of mortality in critically ill patients in the scientific literature. The main-term logistic regression is the parametric algorithm that has been used for constructing both the SAPS-II and APACHE-II scores. This algorithm was included in the Super Learner library so that revised fits of the SAPS-II score based on the current data also competed against other algorithms.

The data used in fitting our prediction algorithm included the 17 variables used in the SAPS-II score: 13 physiological variables (age, Glasgow Coma Scale, systolic blood pressure, heart rate, body temperature, PaO_2/FiO_2 ratio, urinary output, serum urea nitrogen concentration, white blood cell count, serum bicarbonate concentration, sodium concentration, potassium concentration, and bilirubin concentration), type of admission (scheduled surgical, unscheduled surgical, or medical), and three underlying disease variables (acquired immunodeficiency syndrome, metastatic cancer, and haematological cancer derived from ICD-9 discharge codes). We produced two sets of predictions based on the Super Learner; the first based on the 17 variables as they appear in the SAPS-II score (SL1), and the second, on the original, untransformed variables (SL2).

The SICULA prediction algorithm

We refer to the Super Learner-based prediction algorithm using untransformed variables (SL2) as SICULA, an acronym for Super ICU Learning Algorithm. An implementation of the SICULA in JavaScript and R has been made available via a user-friendly web interface. With this web application, clinicians and researchers can obtain the predicted probability of hospital mortality in patients in ICU based on SICULA by inputting patient characteristics.

For the **SICULA** web interface see http://webapps.biostat.berkeley.edu:8080/sicula/

External validation

An external validation of the predictive performance of the SICULA was done with the same metrics but an independent dataset. For external validation, we used data from 200 patients admitted to hospital between Sept 1, 2013, and June 30, 2014. The patients were randomly selected (a random list of patient IDs was generated in all patient IDs found in our local ICU database, and corresponding patients were recruited into our cohort) from the internal anonymous database of patients from the medical, surgical, and trauma ICU at Hôpital Européen Georges Pompidou, Paris, France, a tertiary academic medical centre and level one trauma centre.

Performance measures

A key objective of this study was to compare the predictive performance of scores based on the Super Learner with that of the SAPS-II and SOFA scores. This comparison depended on various measures of predictive performance. First, a mortality prediction algorithm has adequate discrimination if it tends to assign higher severity scores to patients who died in the hospital than to those who did not. We assessed discrimination with the cross-validated area under the receiver-operating characteristic curve (AUROC), reported with corresponding 95% confidence intervals. Discrimination can be graphically shown with the receiver-operating curves (ROC). Additional methods for assessment of discrimination include boxplots of predicted probabilities of death for survivors and non-survivors, and corresponding discrimination slopes, defined as the difference between the mean predicted risks in survivors and non-survivors.

Second, a mortality prediction algorithm is adequately calibrated if predicted and recorded probabilities of death coincide well. We assessed calibration with the Cox calibration test.[12,26,27] Because of its many shortcomings, including poor performance in large samples, we avoided the more conventional Hosmer-Lemeshow statistic.[28,29] Under perfect calibration, a prediction algorithm will satisfy the logistic regression equation:

$$\text{Observed log-odds of death} = \alpha + \beta \times \text{predicted log-odds of death}$$

Where α=0 and β=1. To implement the Cox calibration test, a logistic regression is done to estimate α and β; these estimates suggest the degree of deviation from ideal calibration. The null hypothesis $(\alpha, \beta)=(0,1)$ is tested formally with a U-statistic.[30]

Third, summary reclassification measures, including the continuous Net Reclassification Index (cNRI) and the Integrated Discrimination Improvement (IDI), are relative metrics that have been devised to overcome the limitations of usual discrimination and calibration measures.[31–33] The cNRI comparing severity score A with

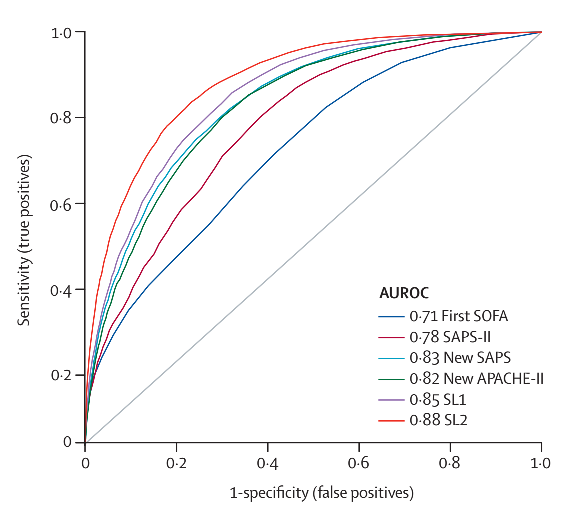

Figure 1: **Receiver-operating characteristics curves**
SL1 with categorised variables and SL2 with non-transformed variables. Results were obtained with 10-fold cross-validation. We also implemented 50-fold cross-validation and noted no material change in the estimated performance of the SICULA algorithm (cross-validated-AUC for the SICULA 0·91 [95% CI 0·90–0·92]). AUROC=area under the receiver-operating characteristics curve. SOFA=Sepsis-related Organ Failure Assessment. SAPS=Simplified Acute Physiology Score. APACHE=Acute Physiology and Chronic Health Evaluation. SL1=Super Learner 1. SL2=Super Learner 2.

score B is defined as twice the difference between the proportion of non-survivors and of survivors, respectively, deemed more severe according to score A rather than score B. The IDI comparing severity score with score B is the average difference in score A between survivors and non-survivors minus the average difference in score B between survivors and non-survivors. Positive values of the cNRI and IDI suggest that score A has better discriminative ability than score B, whereas negative values suggest the opposite. We computed the reclassification tables and associated summary measures to compare each Super Learner proposal with the original SAPS-II score and each of the revised fits of the SAPS-II and APACHE-II scores. All analyses were done with statistical software R (version 2.15.2) for Mac OS X cross-validated AUROC (cv-AUROC),[34] Super Learner,[35] and ROCR.[36]

Role of the funding sources

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

	Mortality prediction
SOFA	0·12 (0·05–0·15)
SAPS-II original version	0·30 (0·08–0·48)
SAPS-II refitted	0·12 (0·03–0·16)
APACHE-II refitted	0·12 (0·03–0·16)
SL1	0·12 (0·02–0·16)
SL2	0·13 (0·01–0·19)

Data are mean (IQR). SOFA=Sepsis-related Organ Failure Assessment. SAPS=Simplified Acute Physiology Score. APACHE=Acute Physiology and Chronic Health Evaluation. SL1=Super Learner 1. SL2=Super Learner 2.

Table 2: **Recorded (3002 [12%]) versus predicted hospital mortality**

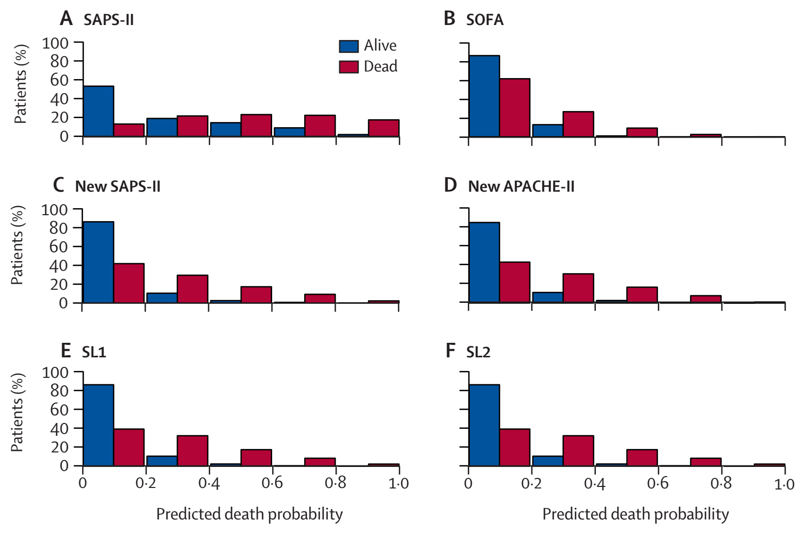

Figure 2: **Distribution of the predicted probability of death in the survivors and the non-survivors**
SOFA=Sepsis-related Organ Failure Assessment. SAPS=Simplified Acute Physiology Score. APACHE=Acute Physiology and Chronic Health Evaluation. SL1=Super Learner 1. SL2=Super Learner 2.

Results

24 508 patients were included in this study. Table 1 shows their baseline characteristics. Figure 1 shows ROCs for hospital mortality prediction. The cv-AUROC was 0·71 (95% CI 0·70–0·72) for the SOFA score, and 0·78 (0·77–0·78) for the SAPS-II score. When refitting the SAPS-II score on our data, the cv-AUROC reached 0·83 (95% CI 0·82–0·83), which is similar to the results obtained with the revised fit of the APACHE-II, which led to an AUROC of 0·82 (0·81–0·83). The two Super Learner (SL1 and SL2) prediction models substantially outperformed the SAPS-II and the SOFA scores, showing a clear advantage of the Super Learner-based prediction algorithms over both the SOFA and SAPS-II scores.

We also investigated discrimination by comparing differences between the predicted probabilities of death in the survivors and the non-survivors with each prediction algorithm (appendix p 3). The discrimination slope was 0·09 for the SOFA score, 0·26 for the SAPS-II score, 0·21 for SL1, and 0·26 for SL2.

Table 2 shows the average predicted probabilities of death based on SL1 and SL2. Probability was similar when we used the SOFA score, the refitted version of the SAPS-II score, and the APACHE-II score. The average probability of death was severely overestimated by the original version of the SAPS-II score (0·30; IQR 0·08–0·48). Figure 2 shows the predicted probabilities of death by survivorship status. Calibration plots suggest a lack of fit for the SAPS-II score (appendix pp 5–7), although the calibration properties were markedly improved by refitting the SAPS-II score. The prediction based on the SOFA and the APACHE-II scores showed excellent calibration properties. For the Super Learner-based predictions, the estimates of α and β were close to the null values. The calibration plots suggest that SL1 is the only method that provides accurate predictions for the entire range of death probability. Indeed, for other algorithms, the predicted probabilities fall close to the ideal calibration line for low probabilities of death but move away from this line as death probabilities increase. For SL1, the predicted probabilities stay close to the ideal calibration line whatever the death probability.

Figure 3 shows the performance of the 12 candidate algorithms, the Discrete Super Learner and the Super Learner combination algorithms, as assessed by CV-MSE and cv-AUROC. As suggested by theory, when either categorised variables (SL1) or untransformed variables (SL2) are used, the Super Learner combination algorithm achieved the same performance as the best of all 12 candidates, with an average CV-MSE of 0·084 (SE 0·001) and an average AUROC of 0·85 (95% CI 0·84–0·85) for SL1 (best algorithm was Bayesian additive regression trees, with CV-MSE 0·084 and AUROC 0·85 [95% CI

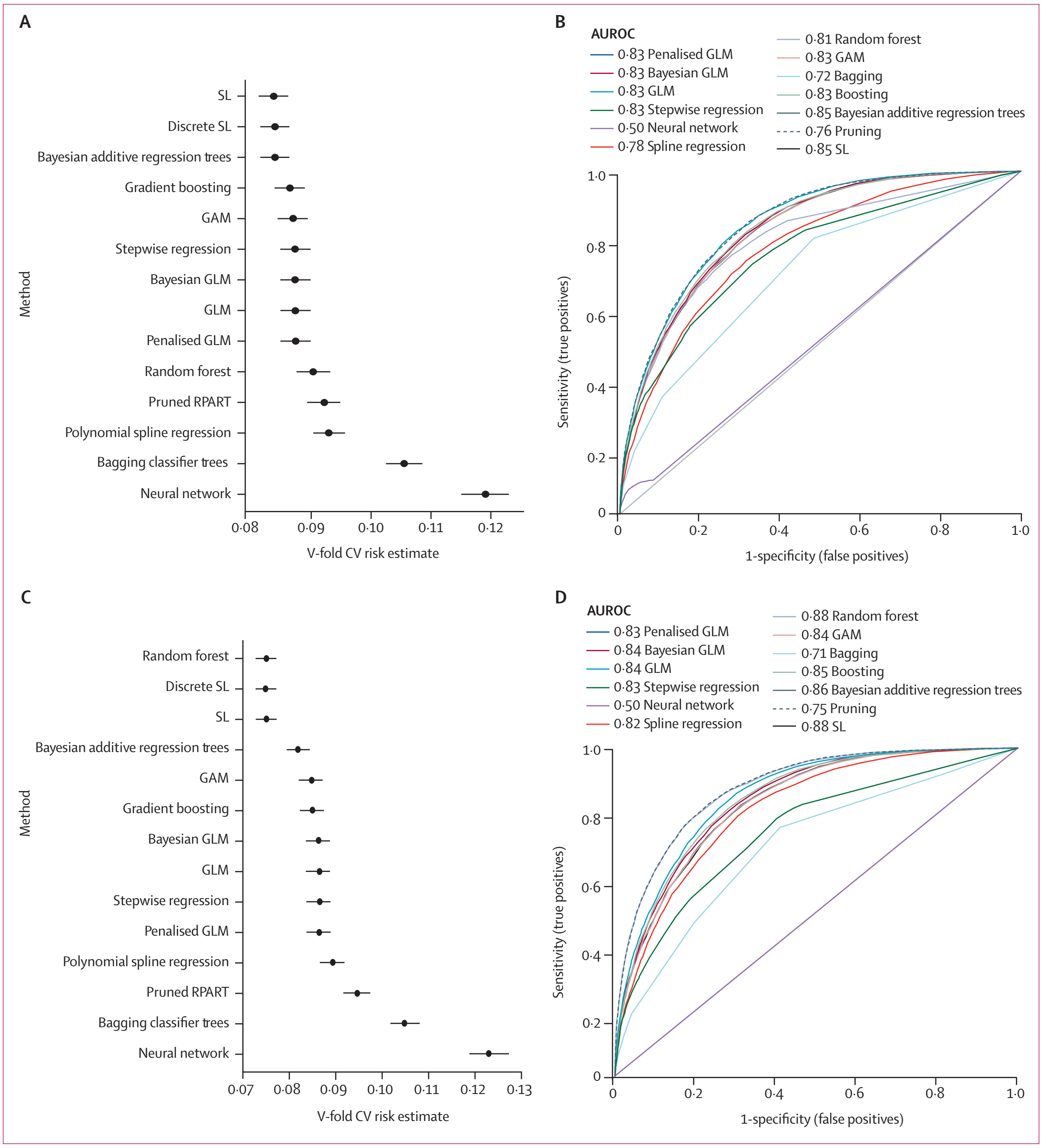

***Figure 3:* Cross-validated mean-squared error for Super Learner and the 12 candidate algorithms included in the library**

(A) SL with categorised variables (Super Learner 1); mean squared error associated with each candidate algorithm. (B) Receiver operating curve (ROC) for each candidate algorithm. (C) Super Learner with non-transformed variables (Super Learner 2); mean squared error associated with each candidate algorithm. (D) ROC for each candidate algorithm. SL=Super Learner. GAM=generalised additive model. GLM=generalised linear model. RPART=recursive partitioning and regression trees. CV=cross-validated. AUROC= area under the receiver-operating characteristics curve.

0·84–0·85]). For the SL2, the average CV-MSE was of 0·076 (SE=0·001) and the average AUROC was 0·88 (95% CI 0·87–0·89; best algorithm was random forests, with CV-MSE 0·076 and AUROC 0·88 [95% CI 0·87–0·89]). In both cases (SL1 and SL2), the Super Learner was better than the main-term logistic regression used to develop the SAPS-II or the APACHE-II score (main-term logistic regression: CV-MSE=0·087 [SE=0·001] and AUROC=0·83 [95% CI 0·82–0·83]).

Table 3 shows reclassification including the SAPS-II score in its original and its actualised versions, the revised APACHE-II score, and the SL1 and SL2 scores. When compared with the classification provided by the original SAPS-II, the actualised SAPS II or the revised

APACHE-II score, the Super Learner-based scores resulted in a downgrade of most patients to a lower risk stratum. We noted this finding especially in patients with a predicted probability of death higher than 0·5. When compared with either the revised SAPS-II or APACHE-II scores, both Super Learner proposals resulted in a large proportion of patients reclassified, especially from higher predicted probability strata to lower ones.

We computed the cNRI and the IDI considering each Super Learner proposal (score A) as the updated model and the original SAPS-II, the new SAPS-II and the new APACHE-II scores (score B) as the initial model. In this case, positive values of the cNRI and IDI would suggest that score A has better discriminative ability than score B, whereas negative values suggest the opposite (table 4).

	Predicted probability according to initial model				Reclassified (%)
	0–0·25	0·25–0·5	0·5–0·75	0·75–1	
SAPS-II, original					
SL1					
0–0·25	13 341	134	3	0	1%
0·25–0·5	4529	723	50	0	86%
0·5–0·75	2703	1090	174	2	96%
0·75–1	444	705	473	137	92%
SL2					
0–0·25	12 932	490	55	1	4%
0·25–0·5	4062	1087	142	11	79%
0·5–0·75	2531	1165	258	15	93%
0·75–1	485	775	448	51	97%
SAPS-II, refitted					
SL1					
0–0·25	20 104	884	30	2	4%
0·25–0·5	894	1426	238	9	44%
0·5–0·75	18	328	361	62	53%
0·75–1	1	14	71	66	57%
SL2					
0–0·25	19 221	1667	124	8	9%
0·25–0·5	765	1478	318	6	42%
0·5–0·75	24	346	367	32	52%
0·75–1	0	26	94	32	79%
APACHE-II, refitted					
SL1					
0–0·25	19 659	1140	107	6	6%
0·25–0·5	1262	1195	296	34	57%
0·5–0·75	89	298	264	71	63%
0·75–1	7	19	33	28	68%
SL2					
0–0·25	18 930	1764	200	18	9%
0·25–0·5	1028	1395	345	19	50%
0·5–0·75	50	333	309	30	57%
0·75–1	2	25	49	11	87%

SL1 with categorised variables. SL2 with non-transformed variables. SAPS=Simplified Acute Physiology Score. SL1=Super Learner 1. SL2=Super Learner 2. APACHE=Acute Physiology and Chronic Health Evaluation.

Table 3: Reclassification

Compared with the original SAPS-II, both the cNRI and IDI were significantly different from zero for SL1. For SL2, the cNRI was significantly different from zero, whereas the IDI was close to zero. Compared with the classification provided by the actualised SAPS II, the cNRI and IDI were significantly different from zero for both SL1 and SL2 (ie, actualised SAPS-II is better). When compared with the actualised APACHE-II score, the cNRI, and IDI were also significantly different from zero (actualised APACHE-II is better) for both SL1 and SL2.

For the patients included in external validation of the SICULA, the main reasons for ICU admission were emergency surgery in 129 patients (65%), elective surgery in 12 patients (6%), and medical (ie, non-surgical) in 59 patients (30%). The median SAPS-II at ICU admission was 40 (18–56). 42 patients (21%) died during their ICU stay. The appendix (pp 8–9) shows ROC curve for SICULA-based hospital mortality prediction. The corresponding AUROC was 0·94 (95% CI 0·90–0·98). The estimated values of α and β were –0·43 and 1·88, respectively (U statistic –0·01, p=0·48), suggesting good calibration properties.

Discussion

The scores developed based on the Super Learner algorithm improved the prediction of hospital mortality in our sample and in an external validation sample, both in terms of discrimination and calibration, compared with the SAPS-II or the APACHE-II scoring systems. The Super Learner severity score (SL2 or SICULA) is based on untransformed versions of the variables used in SAPS-II and APACHE-II, and is available online through a web application. Table 5 shows mortality prediction scores obtained from the SAPS-II, APACHE-II, and SICULA algorithms for three different patient profiles. Specifically, the SAPS-II score is prone to overprediction relative to its two other competitors, except in high-risk surgical patients.

Acknowledging that the assumptions underlying the use of common parametric methods are generally unrealistic in this context for predicting ICU mortality (eg, logistic regression, because the process leading to ICU death is highly complex and therefore unlikely to be adequately captured by a linear relationship with explanatory variables), various investigators have advocated the use of non-parametric techniques for predicting ICU mortality. More than 15 years ago, Dybowski and colleagues[37] assessed neural networks for this purpose and reported a significantly improved AUROC compared with standard logistic regression including second order interactions. However, in a similar setting, Clermont and colleagues[38] later found that logistic regression and neural networks had similar results for ICU mortality prediction. Conflicting results were reported for other non-parametric techniques as well. For instance, Ribas and colleagues[39] reported that use of support vector machines resulted in increased prediction accuracy relative to the APACHE-II score[5] and various shrinkage methods (including the Lasso and ridge regression). Again, these results were tempered when Kim and colleagues[40] reported no clear benefit derived from using neural networks and support vector machines in their sample compared with APACHE-III. Rather, in the latter study, optimum performance was achieved with a decision tree. Similar results have previously been reported with the MIMIC-II dataset.[41] Indeed, a Bayesian ensemble learning algorithm has recently been assessed during an ICU mortality prediction modelling exercise as part of the PhysioNet/Computing in Cardiology Challenge and has shown substantial improvement in prediction performance compared with the SAPS score.[41] During the same study, different authors achieved improved mortality prediction with a method based on support vector machines.[42] Such contradictory results on the relative performance of different prediction methods underscore the fact that no one algorithm invariably outperforms all others.

In any given setting, according to the outcome of interest, the set of explanatory variables available and the underlying population to which it will be applied, the best predictive model might be achieved by a parametric or any of various non-parametric methods. For example, in a situation in which some knowledge about the true shape of the association between the outcome and the explanatory variables is available, a parametric model reflecting this knowledge is likely to outperform any non-parametric technique. The crucial advantage of the Super Learner is that it can include as many candidate algorithms as inputted by investigators, including algorithms that use available scientific knowledge, and in fact borrows strength from diversity in its library. Indeed, established theory suggests that in large samples the Super Learner did at least as well as the (unknown) optimum choice among the library of candidate algorithms.[25] SL1 achieves similar performance as BART, the best candidate when using transformed variables, whereas SL2 achieves similar performance as random forest, which outperformed all other candidates when using untransformed variables (figure 3). Hence, the Super Learner offers a much more flexible alternative to other non-parametric methods.

	SL 1	SL2
SAPS-II, original		
cNRI	0·088 (0·050 to 0·126)	0·247 (0·209 to 0·285)
IDI	–0·048 (–0·055 to –0·041)	–0·001 (–0·010 to –0·008)
SAPS-II, refitted		
cNRI	0·295 (0·257 to 0·333)	0·528 (0·415 to 0·565)
IDI	0·012 (0·008 to 0·017)	0·060 (0·054 to 0·065)
APACHE-II, refitted		
cNRI	0·336 (0·298 to 0·374)	0·561 (0·524 to 0·598)
IDI	0·029 (0·023 to 0·035)	0·076 (0·069 to 0·082)

Data are mean (IQR). SL1 with categorised variables. SL2 with non-transformed variables. SL1=Super Learner 1. SL2=Super Learner 2. SAPS=Simplified Acute Physiology Score. cNRI=continuous Net Reclassification Index. IDI=Integrated Discrimination Improvement. APACHE=Acute Physiology and Chronic Health Evaluation.

Table 4: **Reclassification statistics**

	Patient one: haemorrhagic shock	Patient two: medical sepsis	Patient three: scheduled high-risk surgery
Age (years)	40	80	80
Heart rate (bpm)	120	100	100
Systolic blood pressure (mm Hg)	95	85	100
Glasgow Coma Scale score	8	14	15
Temperature (°C)	36	38	35
Urine output (mL)	700	700	1200
PaO_2/FiO_2	300	200	300
Serum urea (mmol/L)	7	10	7
White blood cell count (10^3/mm^3)	9	19	14
Serum potassium (mmol/L)	4·0	4·8	4·0
Serum sodium (mmol/L)	142	142	142
Serum bicarbonates (μmol/L)	18	18	22
Haematocrit (%)	25%	35%	35%
Bilirubin (μmol/L)	0·8	0·8	0·8
Chronic diseases	None	None	Metastatic cancer
Type of admission	Unscheduled surgery (trauma)	Medical	Scheduled surgery
Mortality prediction			
SAPS-II	46·1%	41·5%	21·3%
APACHE-II	32·2%	23·5%	26·2%
SICULA	29·4%	29·9%	28·7%

SAPS=Simplified Acute Physiology Score. APACHE=Acute Physiology and Chronic Health Evaluation. SICULA=Super ICU Learner Algorithm.

Table 5: **mortality prediction scores obtained from the SAPS-II, APACHE-II, and SICULA algorithms for three different patient profiles**

Our results show that various measures should be considered when assessing the predictive performance of a given severity score (panel). Although the discrepancy between average predicted probability of death and actual recorded in-sample mortality rate was substantial for the original SAPS-II score, it was very small and nearly equal to each of SL1, SL2, the SOFA score and the refitted version of the SAPS-II and APACHE-II scores. However, these findings do not imply that the latter are equally good mortality scores. Indeed, prediction might very well be accurate on average, but still poor at the individual level. Moreover, the accurate average mortality prediction seen with the refitted SAPS-II and APACHE-II scores might be indicative of a certain level of overfitting. A broader assessment of these scores' performance should be considered, namely by carefully studying their discrimination and calibration properties. On one hand, the first SOFA score exhibited very good calibration, yet had very poor discrimination, as shown by the large overlap in predicted probabilities of death between survivors and non-survivors. On the other hand, the

Panel: Research in context

Systematic review

We searched Pubmed and Google Scholar with the following keywords: "ICU", "mortality prediction", "severity scores", "machine learning", "Super Learner", and "non-parametric". No language or date restriction was applied. All appropriate articles were selected based on a careful reading and served as background for our research. Our search showed several attempts had been made to use machine learning techniques in the context of intensive care unit (ICU) mortality prediction, although to the best of our knowledge, none of the reported efforts used ensemble learning techniques. More importantly, the resulting scores are not commonly used or widely available to clinicians and researchers. The most common severity scores in practice date back to the early 1980s and are based on classical logistic regression models.

Interpretation

Our results show that flexible modelling approaches might yield significant improvement in ICU mortality prediction. Our data suggest that instead of relying on one parametric or non-parametric modelling technique, an ensemble machine learning approach should be used to model outcomes as complex as ICU mortality. Clinicians should be aware that prediction based on classical parametric approaches could be misleading. With regards to ICU mortality prediction, the Super ICU Learner Algorithm (SICULA) is a promising alternative that could be valuable both in clinical practice and for research purposes.

SAPS-II score had high discrimination, but was inadequately calibrated in our sample. These results are consistent with previous studies that evaluated the calibration of the SAPS-II score.[15]

The Super Learner offered an appealing tradeoff with good calibration properties and far better discrimination than either the SAPS-II and SOFA scores. Nonetheless, a disclaimer should accompany a criticism of the SOFA score on this basis: in reality, this score was not initially developed for mortality prediction. However, many intensivists use the SOFA score as a surrogate for organ failure quantification and follow-up to assess patients' response to ICU care, and thereby adjust their own perception of likely patient outcomes. For this reason, we chose to assess the performance of SOFA for ICU mortality prediction. In view of the similarity in calibration of the two Super Learner-based scores (SL1 and SL2), we recommend using the Super Learner with untransformed explanatory variables (SL2) in view of its greater discrimination. When considering risk reclassification, the two Super Learner prediction algorithms had similar cNRI, but SL2 clearly had a better IDI. When considering the IDI, the SL1 seemed to perform worse that the SAPS II score. Nonetheless, the IDI should be used carefully because it has similar drawbacks as the AUROC—ie, it summarises prediction characteristics uniformly over all possible classification thresholds even though many of these are unacceptable and would never be considered in practice.[43]

We externally validated the performance of the SICULA with a small dataset obtained from a French ICU. Discrimination performance was excellent. Calibration results were slightly worse than those obtained internally. However, this is mitigated by the fact that the validation sample substantially differed from the training sample, with more severely ill patients, very few patients admitted to hospital for coronary care, and thus a consistently higher hospital mortality rate. Refitting the SICULA with a wider spectrum of ICU patients would probably improve its external validity, which is one of the main goals of the second phase of the SICULA project.

Our study has some limitations. First, we used the SAPS-II and the APACHE-II scores as references although more recent algorithms are available. This was partly because some of the predictors included in the most recent version of these scores were not directly available in the MIMIC-II database. Nonetheless, these scores (eg, SAPS-III and APACHE-III) are associated with the same drawbacks as SAPS-II.[12,15,44] Moreover, those scores are the most widely used scores in practice.[45] Second, our sample comes from one hospital. However, patients in our sample come from five different ICUs, injecting a certain level of heterogeneity in our patient pool. This case-mix heterogeneity might in turn represent a limitation when considering the score for a very specific subpopulation of patients. Moreover, overfitting was mitigated by the use of cross-validation.[46] The patients included in the MIMIC-II cohort seem representative of the overall ICU patient population, as shown by a hospital mortality rate in the MIMIC-II cohort that is similar to the one reported for ICU patients during the same time.[3] Consequently, our score can be expected to show, in other samples, performance characteristics similar to those reported here, at least in samples drawn from similar patient populations. However, by discarding patients with many ICU admissions during the same ICU stay, we might have shrunk the study population toward a less severely ill one. The second phase of the SICULA project will include patients with multiple ICU stays. Additionally, information about do not resuscitate orders or restricted treatments was missing in our dataset and should ideally be taken into account in future work. Third, the large representation in our sample of patients admitted to coronary or cardiac surgery recovery ICU, who often have lower severity scores than medical or surgical ICU patients, might have limited our score's applicability to more critically ill patients. However, further scrutiny showed that the average SAPS-II score in our sample was similar to that reported in similar studies.[15,44]

Of note, results of the discrimination and calibration of the SICULA by ICU type (ie, medical, trauma-surgical, coronary, cardiac surgery recovery, and medicosurgical) showed no substantial difference in prediction performance between units (appendix pp 10–12). Fourth, some variables needed to compute the SAPS-II (eg, elective surgery, underlying disease variables or main reason for ICU admission) were not directly available in the dataset and had to be extrapolated from other data. Finally, a key assumption made was that the poor calibration associated with present severity scores derives from the use of insufficiently flexible statistical models rather than an inappropriate selection of variables included in the model. For this reason and for the sake of providing a fair comparison of our novel score with the SAPS-II score, we included the same explanatory variables as used in SAPS-II. Expansion of the set of explanatory variables used could potentially result in a score with even better predictive performance. In the future, expanding the number of explanatory variables will probably further improve the predictive performance of the score. However, this expansion will probably strengthen further the need for nonparametric approaches and ensemble learning algorithms such as the Super Learner. Indeed, parametric models are known to be less and less adequate as the number of predictors increases.[47] Moreover, when increasing the number of predictors, a sensible trade-off between complexity and performance is even more crucial for the score to still be applicable in practice.

Although additional work remains to be done to validate the resulting prediction algorithm on a large external cohort and to incorporate additional predictor variables, an accessible, user-friendly web implementation of our scoring procedure has been made available. This implementation allows clinicians to use our score in their own practice, say as an aid in working out treatment allocation, provides an opportunity for clinician-researchers to validate our algorithm within the context of their own patient populations, and serves as an improved risk stratification method for use in clinical research. This is in rather sharp contrast with other instances in which scores have been developed using complex machine learning methods but the resulting scores cannot be readily calculated by clinicians. Indeed, we found no example in which an implementation of a published scoring procedure was made publicly available on the web. In addition, we have made the corresponding R code available to other investigators in an online appendix.

We conclude from this first stage of the SICULA project that, in this population, the prediction of hospital mortality based on the SICULA prediction algorithm achieves significantly improved performance, both in terms of calibration and discrimination, compared with conventional severity scores. The SICULA prediction algorithm is a promising alternative that could be valuable both in clinical practice and for research purposes. External validation of results of this study in different populations, especially outside of the USA, providing periodic updates of the SICULA fit, and assessment of the potential benefit of including additional variables in the score remain important future challenges that will be tackled in the second stage of the SICULA project. Before an unequivocal recommendation of the widespread use of our algorithm can be made, our findings need to be confirmed in this second phase. Nevertheless, we believe the currently available web implementation of SICULA (appendix p 13) should prove useful to both clinicians and other investigators in critical care medicine.

Contributors

RP contributed to the concept of the study, analysis, and interpretation of data, and drafting and revision of the report. MLP and MC contributed to the analysis and interpretation of data, and drafting and revision of the report. MRR contributed to the analysis of data and revision of the report; SC and MJvdL contributed to the interpretation of data and revision of the report.

Declaration of interests

RP reports grants from Assistance Publique Hôpitaux de Paris and grants from the Fulbright Foundation, during the conduct of the study.
MLP reports grants from the Doris Duke Clinical Scientist Development Award, during the conduct of the study. MJvdL reports grants from the NIH, during the conduct of the study. All other authors declare no competing interests.

Acknowledgements

This study was funded by the Fulbright Foundation, Assistance Publique–Hôpitaux de Paris (RP), Doris Duke Clinical Scientist Development Award (MP), and an NIH grant (2R01AI074345-06A1; MvdL).

References

1 Halpern NA, Pastores SM. Critical care medicine in the United States 2000–2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Crit Care Med* 2010; **38:** 65–71.

2 Ridley S, Morris S. Cost effectiveness of adult intensive care in the UK. *Anaesthesia* 2007; **62:** 547–54.

3 Zimmerman JE, Kramer AA, Knaus WA. Changes in hospital mortality for United States intensive care unit admissions from 1988 to 2012. *Crit Care* 2013; **17:** R81.

4 Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med* 1981; **9:** 591–97.

5 Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; **13:** 818–29.

6 Le Gall JR, Loirat P, Alperovitch A, et al. A simplified acute physiology score for ICU patients. *Crit Care Med* 1984; **12:** 975–77.

7 Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993; **270:** 2957–63.

8 Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *JAMA* 1993; **270:** 2478–86.

9 Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991; **100:** 1619–36.

10 Le Gall JR, Neumann A, Hemery F, et al. Mortality prediction using SAPS II: an update for French intensive care units. *Crit Care* 2005; **9:** R645–52.

11 Nassar AP Jr, Mocelin AO, Nunes ALB, et al. Caution when using prognostic models: a prospective comparison of 3 recent prognostic models. *J Crit Care* 2012; **27**: 423.e1–423.e7.

12 Poole D, Rossi C, Latronico N, Rossi G, Finazzi S, Bertolini G, and the GiViTI. Comparison between SAPS II and SAPS 3 in predicting hospital mortality in a cohort of 103 Italian ICUs. Is new always better? *Intensive Care Med* 2012; **38:** 1280–88.

13 Metnitz B, Schaden E, Moreno R, Le Gall J-R, Bauer P, Metnitz PGH, and the ASDI Study Group. Austrian validation and customization of the SAPS 3 Admission Score. *Intensive Care Med* 2009; **35:** 616–22.

14 Moreno RP, Metnitz PGH, Almeida E, et al, and the SAPS 3 Investigators. SAPS 3--From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* 2005; **31:** 1345–55.

15 Beck DH, Smith GB, Pappachan JV, Millar B. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. *Intensive Care Med* 2003; **29:** 249–56.

16 Aegerter P, Boumendil A, Retbi A, Minvielle E, Dervaux B, Guidet B. SAPS II revisited. *Intensive Care Med* 2005; **31:** 416–23.

17 Ledoux D, Canivet J-L, Preiser J-C, Lefrancq J, Damas P. SAPS 3 admission score: an external validation in a general intensive care population. *Intensive Care Med* 2008; **34:** 1873–77.

18 Lee J, Scott DJ, Villarroel M, Clifford GD, Saeed M, Mark RG. Open-access MIMIC-II database for intensive care research. *Conf Proc IEEE Eng Med Biol Soc* 2011; **2011:** 8315–18.

19 Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. *Crit Care Med* 2011; **39:** 952–60.

20 Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 2000; **101:** E215–20.

21 Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med* 1996; **22:** 707–10.

22 Neamatullah I, Douglass MM, Lehman LW, et al. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak* 2008; **8:** 32–48.

23 Dudoit S, Van Der Laan MJ. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat Methodol* 2003; **2:** 131–54.

24 Van Der Laan MJ, Dudoit S. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. *UC Berkeley Division of Biostatistics Working Paper Series* 2003; **Working Paper**: 1–103.

25 Van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol* 2007; **6**: 25.

26 Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958; **45:** 562–65.

27 Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. *Crit Care Med* 2006; **34:** 1378–88.

28 Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit Care Med* 2007; **35:** 2052–56.

29 Bertolini G, D'Amico R, Nardi D, Tinazzi A, Apolone G. One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *J Epidemiol Biostat* 2000; **5:** 251–53.

30 Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Stat Med* 1991; **10:** 1213–26.

31 Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; **115:** 928–35.

32 Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem* 2008; **54:** 17–23.

33 Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008; **27:** 157–72, discussion 207–12.

34 LeDell E, Petersen M, van der Laan M, LeDell ME. Package 'cvAUC'. ftp://ftp.sam.math.ethz.ch/sfs/Software/R-CRAN/web/packages/cvAUC/cvAUC.pdf (accessed March 13, 2014).

35 Polley E, van der Laan M. SuperLearner: Super Learner Prediction 2014. http://CRAN.R-project.org/package=SuperLearner (accessed March 13, 2014).

36 Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics* 2005; **21:** 3940–41.

37 Dybowski R, Weller P, Chang R, Gant V. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. *Lancet* 1996; **347:** 1146–50.

38 Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Crit Care Med* 2001; **29:** 291–96.

39 Ribas VJ, López JC, Ruiz-Sanmartin A, et al. Severe sepsis mortality prediction with relevance vector machines. *Conf Proc IEEE Eng Med Biol Soc* 2011; **2011:** 100–03.

40 Kim S, Kim W, Park RW. A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. *Healthc Inform Res* 2011; **17:** 232–43.

41 Johnson A, Dunkley N, Mayaud L, Tsanas A, Kramer A, Clifford G. Patient-specific Predictions in the ICU using a Bayesian Ensemble. *Comput Cardiol* 2012; **2012:** 249–52.

42 Citi L, Barbieri R. PhysioNet 2012 challenge: predicting mortality of ICU patients using a cascaded SVM-GLM paradigm. *Comput Cardiol* 2012; **39:** 257–60.

43 Greenland S. The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929). *Stat Med* 2008; **27:** 199–206.

44 Sakr Y, Krauss C, Amaral ACKB, et al. Comparison of the performance of SAPS II, SAPS 3, APACHE II, and their customized prognostic models in a surgical intensive care unit. *Br J Anaesth* 2008; **101:** 798–803.

45 Rosenberg AL. Recent innovations in intensive care unit risk-prediction models. *Curr Opin Crit Care* 2002; **8:** 321–30.

46 Steyerberg EW, Harrell FE Jr, Borsboom GJJ, Eijkemans MJ, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001; **54:** 774–81.

47 Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods* 2009; **14:** 323–48.