Chapter 20
Mortality Prediction in the ICU Based on MIMIC-II Results from the Super ICU Learner Algorithm (SICULA) Project

Romain Pirracchio

Learning Objectives
In this chapter, we illustrate the use of MIMIC II clinical data, non-parametric prediction algorithm, ensemble machine learning, and the Super Learner algorithm.

20.1 Introduction

Predicting mortality in patients hospitalized in intensive care units (ICU) is crucial for assessing severity of illness and adjudicating the value of novel treatments, interventions and health care policies. Several severity scores have been developed with the objective of predicting hospital mortality from baseline patient characteristics, defined as measurements obtained within the first 24 h after ICU admission. The first scores proposed, APACHE [1] (Acute Physiology and Chronic Health Evaluation), APACHE II [2], and SAPS [3] (Simplified Acute Physiology Score), relied upon subjective methods for variable importance measure, namely by prompting 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 [4] were subsequently developed using statistical modeling techniques [4–7]. To this day, the SAPS II [4] and APACHE II [2] scores remain the most widely used in clinical practice. However, since first being published, they have been modified several times in order to improve their predictive performance [6–11]. Despite these extensions of SAPS, predicted hospital mortality remains generally overestimated [8, 9, 12–14]. As an illustration, Poole et al. [9] compared the SAPS II and the SAPS3 performance in a cohort of more than 28,000 admissions to 10 different Italian ICUs. They concluded that both scores provided unreliable predictions, but unexpectedly the newer SAPS 3 turned out to overpredict mortality more than the

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_20

older SAPS II. Consistently, Nassar et al. [8] assessed the performance of the APACHE IV, the SAPS 3 and the Mortality Probability Model III [MPM(0)-III] in a population admitted at 3 medical-surgical Brazilian intensive care units and found that all models showed poor calibration, while discrimination was very good for all of them.

Most ICU severity scores rely on a logistic regression model. Such models impose stringent constraints on the relationship between explanatory variables and risk of death. For instance, main term logistic regression relies on the assumption of a linear and additive relationship between the outcome and its predictors. Given the complexity of the processes underlying death in ICU patients, this assumption might be unrealistic.

Given that the true relationship between risk of mortality in the ICU and explanatory variables is unknown, we expect that prediction can be improved by using an automated nonparametric algorithm to estimate risk of death without requiring any specification about the shape of the underlying relationship. Indeed, nonparametric algorithms offer the great advantage of not relying on any assumption about the underlying distribution, which make them more suited to fit such complex data. Some studies have evaluated the benefit of nonparametric approaches, namely based on neural networks or data-mining, to predict hospital mortality in ICU patients [15–20]. These studies unanimously concluded that nonparametric methods might perform at least as well as standard logistic regression in predicting ICU mortality.

Recently, the *Super Learner* was developed as a nonparametric technique for selecting an optimal regression algorithm among a given set of candidate algorithms provided by the user [21]. The *Super Learner* ranks the algorithms according to their prediction performance, and then builds an aggregate algorithm obtained as the optimal weighted combination of the candidate algorithms. Theoretical results have demonstrated that the *Super Learner* performs no worse than the optimal choice among the provided library of candidate algorithms, at least in large samples. It capitalizes on the richness of the library it builds upon and generally offers gains over any specific candidate algorithm in terms of flexibility to accurately fit the data.

The primary aim of this study was to develop a scoring procedure for ICU patients based on the *Super Learner* using data from the Medical Information Mart for Intensive Care II (MIMIC-II) study [22–24], and to determine whether it results in improved mortality prediction relative to the SAPS II, the APACHE II and the SOFA scores. Complete results of this study have been published in 2015 in the Lancet Respiratory Medicine [25]. We also wished to develop an easily-accessible user-friendly web implementation of our scoring procedure, even despite the complexity of our approach (http://webapps.biostat.berkeley.edu:8080/sicula/).

20.2 Dataset and Pre-preprocessing

20.2.1 Data Collection and Patients Characteristics

The MIMIC-II study [22–24] includes all patients admitted to an ICU at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA since 2001. For the sake of the present study, only data from MIMIC-II version 26 (2001–2008) on adult ICU patients were included. Patients younger than 16 years were not included. For patients with multiple admission, we only considered the first ICU stay. A total of 24,508 patients were included in this study.

20.2.2 Patient Inclusion and Measures

Two categories of data were collected: clinical data, aggregated from ICU information systems and hospital archives, and high-resolution physiologic data (waveforms and time series of derived physiologic measurements), recorded on bedside monitors. Clinical data were obtained from the CareVue Clinical Information System (Philips Healthcare, Andover, Massachusetts) deployed in all study ICUs, and from hospital electronic archives. The data included time-stamped nurse-verified physiologic measurements (e.g., hourly documentation of heart rate, arterial blood pressure, pulmonary artery pressure), nurses' and respiratory therapists' progress notes, continuous intravenous (IV) drip medications, fluid balances, patient demographics, interpretations of imaging studies, physician orders, discharge summaries, and ICD-9 codes. Comprehensive diagnostic laboratory results (e.g., blood chemistry, complete blood counts, arterial blood gases, 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 [4] and SOFA scores [26].

We first took an inventory of all available recorded characteristics required to evaluate the different scores considered. Raw data from the MIMIC II database version 26 were then extracted. We decided to use only R functions (without any SQL routines) as most of our researchers only have R package knowledge. Each table within each patient datafile were checked for the different characteristics and extracted. Finally, we created a global CSV file including all data and easily manipulable with R.

Baseline variables and outcomes are summarized in Table 20.1.

Table 20.1 Baseline characteristics and outcome measures

	Overall population (n = 24,508)	Dead at hospital discharge (n = 3002)	Alive at hospital discharge (n = 21,506)
Age	65 [51–77]	74 [59–83]	64 [50–76]
Gender (female)	13,838 (56.5 %)	1607 (53.5 %)	12,231 (56.9 %)
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]
Origin			
Medical	2453 (10 %)	240 (8 %)	2213 (10.3 %)
Trauma	7703 (31.4 %)	1055 (35.1 %)	6648 (30.9 %)
Emergency surgery	10,803 (44.1 %)	1583 (52.7 %)	9220 (42.9 %)
Scheduled surgery	3549 (14.5 %)	124 (4.1 %)	3425 (15.9 %)
Site			
MICU	7488 (30.6 %)	1265 (42.1 %)	6223 (28.9 %)
MSICU	2686 (11 %)	347 (11.6 %)	2339 (10.9 %)
CCU	5285 (21.6 %)	633 (21.1 %)	4652 (21.6 %)
CSRU	8100 (33.1 %)	664 (22.1 %)	7436 (34.6 %)
TSICU	949 (3.9 %)	93 (3.1 %)	856 (4 %)
HR (bpm)	87 [75–100]	92 [78–109]	86 [75–99]
MAP (mmHg)	81 [70–94]	78 [65–94]	82 [71–94]
RR (cpm)	14 [12–20]	18 [14–23]	14 [12–18]
Na (mmol/l)	139 [136–141]	138 [135–141]	139 [136–141]
K (mmol/l)	4.2 [3.8–4.6]	4.2 [3.8–4.8]	4.2 [3.8–4.6]
HCO_3 (mmol/l)	26 [22–28]	24 [20–28]	26 [23–28]
WBC ($10^3/mm^3$)	10.3 [7.5–14.4]	11.6 [7.9–16.9]	10.2 [7.4–14.1]
P/F ratio	281 [130–447]	174 [90–352]	312 [145–461]
Ht (%)	34.7 [30.4–39]	33.8 [29.8–38]	34.8 [30.5–39.1]
Urea (mmol/l)	20 [14–31]	28 [18–46]	19 [13–29]
Bilirubine (mg/dl)	0.6 [0.4–1]	0.7 [0.4–1.5]	0.6 [0.4–0.9]
Hospital LOS (days)	8 [4–14]	9 [4–17]	8 [4–14]
ICU death (%)	1978 (8.1 %)	1978 (65.9 %)	–
Hospital death (%)	3002 (12.2 %)	–	–

Continuous variables are presented as median [InterQuartile Range]; binary or categorical variables as count (%)

20.3 Methods

20.3.1 Prediction Algorithms

The primary outcome measure was hospital mortality. A total of 1978 deaths occurred in ICU (estimated mortality rate: 8.1 %, 95 %CI: 7.7–8.4), and 1024 additional deaths were observed after ICU discharge, resulting in an estimated hospital mortality rate of 12.2 % (95 %CI: 11.8–12.7).

The data recorded within the first 24 h following ICU admission were used to compute two of the most widely used severity scores, namely the SAPS II [4] and SOFA [26] scores. Individual mortality prediction for the SAPS II score was calculated as defined by its authors [4]:

$$\log\left[\frac{\text{pr(death)}}{1-\text{pr(death)}}\right] = -7.7631 + 0.0737 * \text{SAPSII} + 0.9971 * \log(1 + \text{SAPSII})$$

In addition, we developed a new version of the SAPS II score, by fitting to our data a main-term logistic regression model using the same explanatory variables as those used in the original SAPS II score [4]: age, heart rate, systolic blood pressure, body temperature Glasgow Coma Scale, mechanical ventilation, PaO_2, FiO_2, urine output, BUN (blood urea nitrogen), blood sodium, potassium, bicarbonates, bilirubin, white blood cells, chronic disease (AIDS, metastatic cancer, hematologic malignancy) and type of admission (elective surgery, medical, unscheduled surgery). The same procedure was used to build a new version of the APACHE II score [2]. Finally, because the SOFA score [26] is widely used in clinical practice as a proxy for outcome prediction, it was also computed for all subjects. Mortality prediction based on the SOFA score was obtained by regressing hospital mortality on the SOFA score using a main-term logistic regression. These two algorithms for mortality prediction were compared to our *Super Learner*-based proposal.

The *Super Learner* has been proposed as a method for selecting via cross-validation the optimal regression algorithm among all weighted combinations of a set of given candidate algorithms, henceforth referred to as the library [21, 27, 28] (Fig. 20.1). To implement the *Super Learner*, a user must provide a customized collection of various data-fitting algorithms. The *Super Learner* then estimates the risk associated to each algorithm in the provided collection using cross-validation. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the *training set*), and validating the analysis on the other subset (called the *validation set* or *testing set*). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds. From this estimation of the risk associated with each candidate algorithm, the *Super Learner* builds an aggregate algorithm obtained as the optimal weighted combination of the candidate algorithms. Theoretical results suggest that to optimize the performance of the

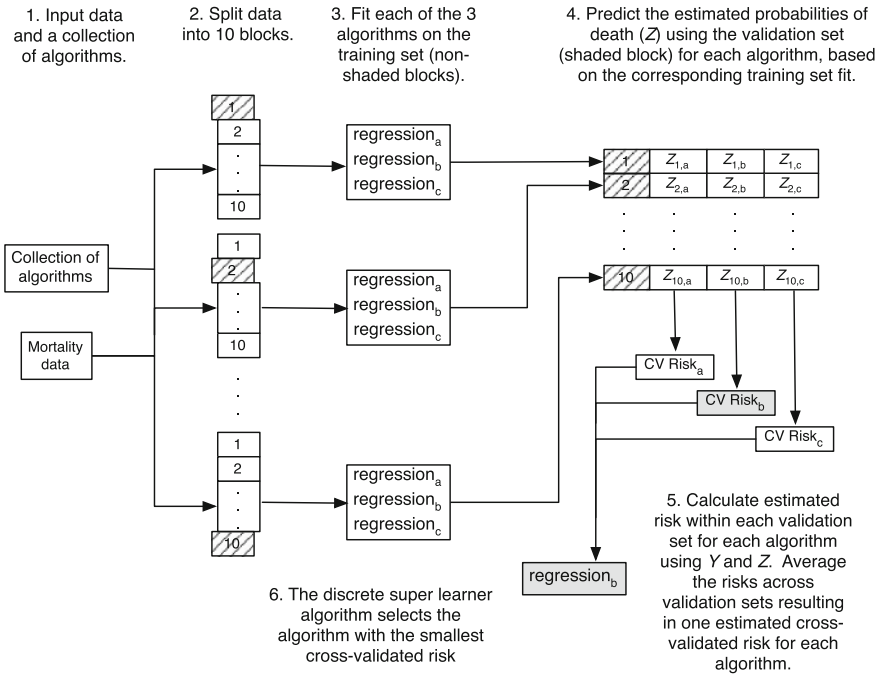

Fig. 20.1 Super learner algorithm. From van der Laan, targeted learning 2011 (with permission) [41]

resulting algorithm, the inputted library should include as many sensible algorithms as possible.

In this study, the library size was limited to 12 algorithms (list available in the Appendix) for computational reasons. Among these 12 algorithms, some were parametric such as logistic regression of affiliated methods classically used for ICU scoring systems, and some non-parametric i.e. methods that fit the data without any assumption concerning the underlying data distribution. In the present study, we chose the library to include most of parametric (including regression models with various combinations of main and interaction terms as well as splines, and fitted using maximum likelihood with or without penalization) and nonparametric algorithm, previously evaluated for the prediction of mortality in critically ill patients in the 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 SL library so that revised fits of the SAPS II score based on the current data also competed against other algorithms.

Comparison of the 12 algorithms relied on 10-fold cross-validation. The data are first split into 10 mutually exclusive and exhaustive blocks of approximately equal size. Each algorithm is fitted on a the 9 blocks corresponding to the training set and then this fit used to predict mortality for all patients in the remaining block used a

validation set. The squared errors between predicted and observed outcomes are averaged. The performance of each algorithm is evaluated in this manner. This procedure is repeated exactly 10 times, with a different block used as validation set every time. Performance measures are aggregated over all 10 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 not a single 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, as overfitting is more likely to occur when training and validation sets intersect.

Candidate algorithms were ranked according to their CV-MSE and the algorithm with least CV-MSE was identified. This algorithm was then refitted using all available data, leading to a prediction rule referred to as the *Discrete Super Learner*. Subsequently, the prediction rule consisting of the CV-MSE-minimizing weighted convex combination of all candidate algorithms was also computed and refitted on all data. This is what we refer to as the *Super Learner* combination algorithm [28].

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 level, white blood cells count, serum bicarbonate level, sodium level, potassium level and bilirubin level), type of admission (scheduled surgical, unscheduled surgical, or medical), and three underlying disease variables (acquired immunodeficiency syndrome, metastatic cancer, and hematologic malignancy derived from ICD-9 discharge codes). Two sets of predictions based on the *Super Learner* were produced: 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).

20.3.2 Performance Metrics

A key objective of this study was to compare the predictive performance of scores based on the *Super Learner* to that of the SAPS II and SOFA scores. This comparison hinged on a variety of measures of predictive performance, described below.

1. A mortality prediction algorithm is said to have adequate discrimination if it tends to assign higher severity scores to patients that died in the hospital compared to those that did not. We evaluated discrimination using the cross-validated area under the receiver-operating characteristic curve (AUROC), reported with corresponding 95 % confidence interval (95 % CI). Discrimination can be graphically illustrated using the receiver-operating (ROC) curves. Additional tools for assessing 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. All these are provided below.
2. A mortality prediction algorithm is said to be adequately calibrated if predicted and observed probabilities of death coincide rather well. We assessed calibration using the Cox calibration test [9, 29, 30]. Because of its numerous shortcoming, including poor performance in large samples, the more conventional Hosmer-Lemeshow statistic was avoided [31, 32]. Under perfect calibration, a prediction algorithm will satisfy the logistic regression equation 'observed log-odds of death = $\alpha + \beta^*$ predicted log-odds of death' with $\alpha = 0$. To implement the Cox calibration test, a logistic regression is performed 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 using a U-statistic [33].
3. Summary reclassification measures, including the Continuous Net Reclassification Index (cNRI) and the Integrated Discrimination Improvement (IDI), are relative metrics which have been devised to overcome the limitations of usual discrimination and calibration measures [34–36]. The cNRI comparing severity score A to 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 A to 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 indicate that score A has better discriminative ability than score B, whereas negative values indicate the opposite. We computed the reclassification tables and associated summary measures to compare each *Super Learner* proposal to the original SAPS II score and each of the revised fits of the SAPS II and APACHE II scores.

All analyses were performed using statistical software R version 2.15.2 for Mac OS X (The R Foundation for Statistical Computing, Vienna, Austria; specific packages: cvAUC, Super Learner and ROCR). Relevant R codes are provided in Appendix.

20.4 Analysis

20.4.1 Discrimination

The ROC curves for hospital mortality prediction are provided below (Fig. 20.2). The cross-validated AUROC was 0.71 (95 %CI: 0.70–0.72) for the SOFA score, and 0.78 (95 %CI: 0.77–0.78) for the SAPS II score. When refitting the SAPS II score on our data, the AUROC reached 0.83 (95 %CI: 0.82–0.83); this is similar to the results obtained with the revised fit of the APACHE II, which led to an AUROC of 0.82 (95 %CI: 0.81–0.83). The two *Super Learner* (SL1 and SL2) prediction models substantially outperformed the SAPS II and the SOFA score. The AUROC

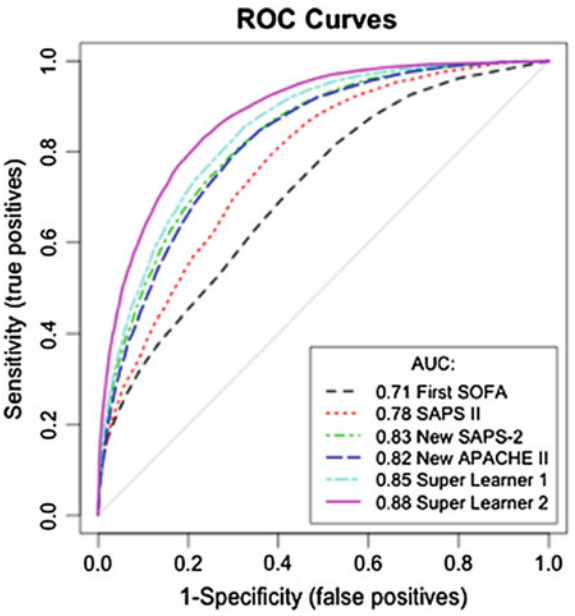

Fig. 20.2 Receiver-operating characteristics curves. Super learner 1: super learner with categorized variables; super learner 2: super learner with non-transformed variables

was 0.85 (95 %CI: 0.84–0.85) for SL1, and 0.88 (95 %CI: 0.87–0.89) for SL2, revealing a clear advantage of the Super Learner-based prediction algorithms over both the SOFA and SAPS II scores.

Discrimination was also evaluated by comparing differences between the predicted probabilities of death among the survivors and the non-survivors using each prediction algorithm. The discrimination slope equaled 0.09 for the SOFA score, 0.26 for the SAPS II score, 0.21 for SL1, and 0.26 for SL2.

20.4.2 Calibration

Calibration plots (Fig. 20.3) indicate a lack of fit for the SAPS II score. The estimated values of α and β were of −1.51 and 0.72 respectively (U statistic = 0.25, $p < 0.0001$). The calibration properties were markedly improved by refitting the SAPS II score: $\alpha < 0.0001$ and $\beta = 1$ ($U < 0.0001$, $p = 1.00$). The prediction based on the SOFA and the APACHE II scores exhibited excellent calibration properties, as reflected by $\alpha < 0.0001$ and $\beta = 1$ ($U < 0.0001$, $p = 1.00$). For the Super Learner-based predictions, despite U-statistics significantly different from zero, the estimates of α and β were close to the null values: SL1: 0.14 and 1.04, respectively ($U = 0.0007$, $p = 0.0001$); SL2: 0.24 and 1.25, respectively ($U = 0.006$, $p < 0.0001$).

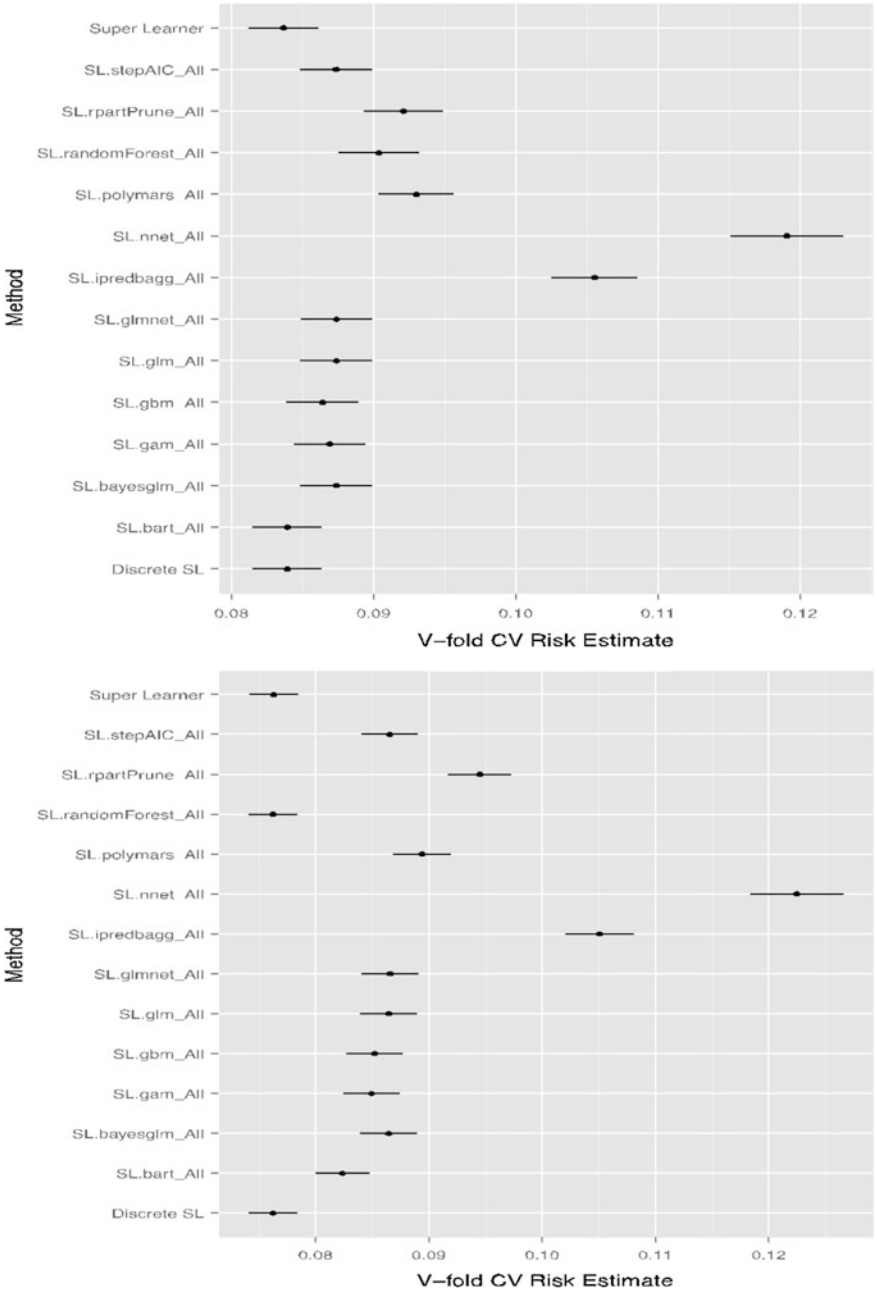

Fig. 20.3 Calibration and discrimination plots for SAPS 2 (*upper panel*) and SL1 (*lower panel*)

20.4.3 Super Learner Library

The performance of the 12 candidate algorithms, the Discrete *Super Learner* and the *Super Learner* combination algorithms, as evaluated by CV-MSE and CV-AUROC, are illustrated in Fig. 20.4.

As suggested by theory, when using either categorized variables (SL1) or untransformed variables (SL2), 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 single algorithm: Bayesian Additive Regression Trees, with CV-MSE = 0.084 and AUROC = 0.84 (95 %CI: 0.84, 0.85)]. For the SL2, the average CV-MSE was of 0.076 (SE = 0.001) and the average AUROC of 0.88 (95 %CI: 0.87–0.89) [best single algorithm: 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* outperformed 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)].

20.4.4 Reclassification Tables

The reclassification *tables involving the SAPS* II score in its original and its actualized versions, the revised APACHE II score, and the SL1 and SL2 scores are provided in Table 20.2. When compared to the classification provided by the original SAPS II, the actualized SAPS II or the revised APACHE II score, the Super Learner-based scores resulted in a downgrade of a large majority of patients to a lower risk stratum. This was especially the case for patients with a predicted probability of death above 0.5.

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 indicate that score A has better discriminative ability than score B, whereas negative values indicate the opposite. For SL1, both the cNRI (cNRI = 0.088 (95 %CI: 0.050, 0.126), $p < 0.0001$) and IDI (IDI = −0.048 (95 % CI: −0.055, −0.041), $p < 0.0001$) were significantly different from zero. For SL2, the cNRI was significantly different from zero (cNRI = 0.247 (95 %CI: 0.209, 0.285), $p < 0.0001$), while the IDI was close to zero (IDI = −0.001 (95 %CI: −0.010, −0.008), $p = 0.80$). When compared to the classification provided by the actualized SAPS II, the cNRI and IDI were significantly different from zero for both SL1 and SL2: cNRI = 0.295 (95 %CI: 0.257, 0.333), $p < 0.0001$ and IDI = 0.012 (95 %CI: 0.008, 0.017), $p < 0.0001$ for SL1; cNRI = 0.528 (95 %CI: 0.415, 0.565), $p < 0.0001$ and IDI = 0.060 (95 %CI: 0.054, 0.065), $p < 0.0001$ for SL2. When compared to the actualized APACHE II score, the cNRI and IDI were also

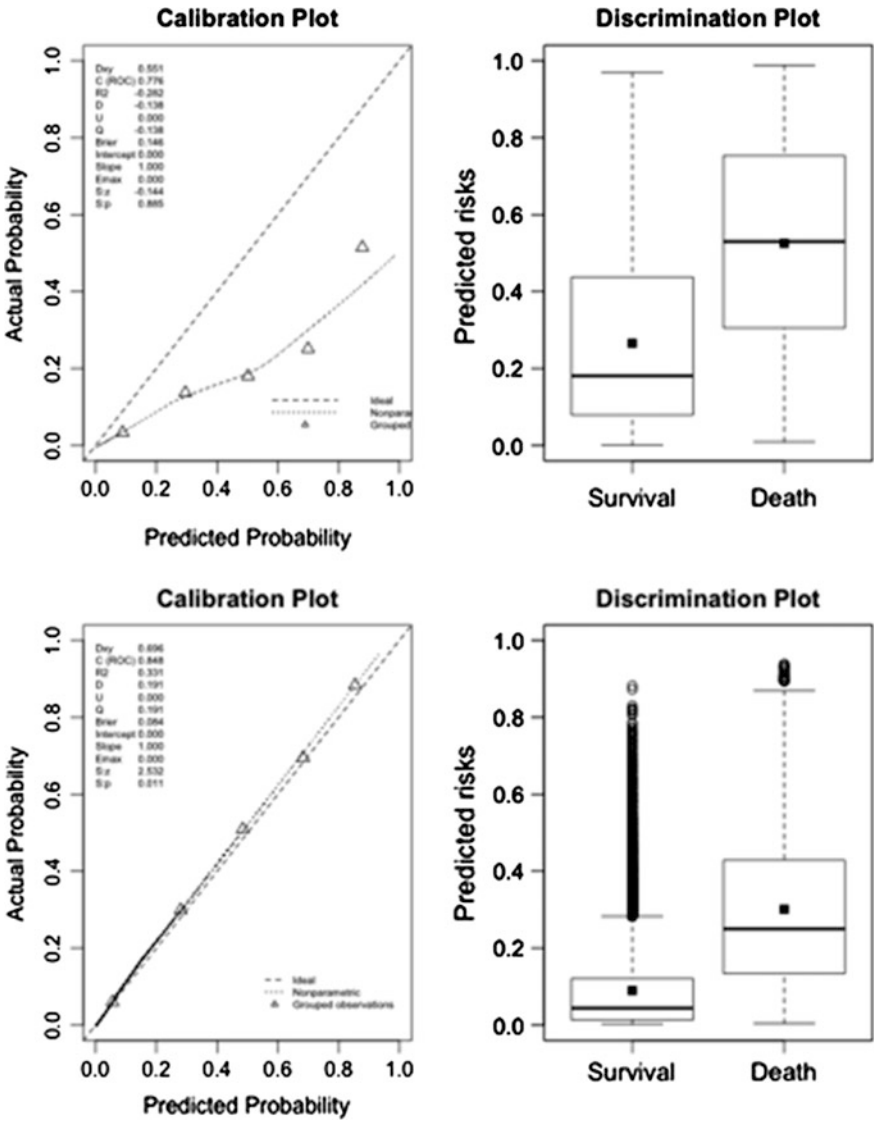

Fig. 20.4 Cross-validated mean-squared error for the super learner and the 12 candidate algorithms included in the library. Upper panel concerns the super learner with categorized variables (super learner 1): mean squared error (MSE) associated with each candidate algorithm (*top figure*)—receiver operating curves (ROC) for each candidate algorithm (*bottom figure*); lower panel concerns the super learner with non-transformed variables (super learner 2): mean squared error (MSE) associated with each candidate algorithm (*top figure*)—receiver operating curves (ROC) for each candidate algorithm (*bottom figure*)

Table 20.2 Reclassification tables

	Updated model				
	0–0.25	0.25–0.5	0.5–0.75	0.75–1	% Reclassified
Super learner 1					
Initial model: original SAPS II					
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 %
Super learner 2					
Initial model: original SAPS II					
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 %
Super learner 1					
Initial model: new SAPS II					
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 %
Super learner 2					
Initial model: new SAPS II					
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 %
Super learner 1					
Initial model: new APACHE II					
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 %
Super learner 2					
Initial model: new APACHE II					
0–0.25	18,930	1764	200	18	9 %
0.25–0.5	1028	1395	345	19	50 %

(continued)

Table 20.2 (continued)

	Updated model				
	0–0.25	0.25–0.5	0.5–0.75	0.75–1	% Reclassified
0.5–0.75	50	333	309	30	57 %
0.75–1	2	25	49	11	87 %

Super learner 1: super learner with categorized variables; super learner 2: super learner with non-transformed variables

significantly different from zero for both SL1 and SL2: cNRI = 0.336 (95 %CI: 0.298, 0.374), $p < 0.0001$ and IDI = 0.029 (95 %CI: 0.023, 0.035), $p < 0.0001$ for SL1; cNRI = 0.561 (95 %CI: 0.524, 0.598), $p < 0.0001$ and IDI = 0.076 (95 %CI: 0.069, 0.082) for SL2. When compared either to the new SAPS II or the new APACHE II score, both Super Learner proposals resulted in a large proportion of patients reclassified, especially from high predicted probability strata to lower ones.

20.5 Discussion

The new scores based on the *Super Learner* improve the prediction of hospital mortality in this sample, both in terms of discrimination and calibration, as compared to the SAPS II or the APACHE II scoring systems. The Super Learner severity score based on untransformed variables, also referred to as SL2 or SICULA, is available online through a web application. An ancillary important result is that the MIMIC-II database can easily and reliably serve to develop new severity score for ICU patients.

Our results illustrate the crucial advantage of the Super Learner that can include as many candidate algorithms as inputted by investigators, including algorithms reflecting available scientific knowledge, and in fact borrows strength from diversity in its library. Indeed, established theory indicates that in large samples the *Super Learner* performs at least as well as the (unknown) optimal choice among the library of candidate algorithms [28]. This is illustrated by comparing the CV-MSE associated with each algorithm included in the library: SL1 achieves similar performance as BART, which is the best candidate in the case, while SL2 achieves similar performance as random forest, which outperformed all other candidates in this case. Hence, the *Super Learner* offers a more flexible alternative to other nonparametric methods.

Given 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. It should be emphasized that, when considering the IDI, the SL1 seemed to perform worse that the SAPS II score. Nonetheless, the IDI must be used carefully since it suffers from similar drawbacks as the AUROC: it

summarizes prediction characteristics uniformly over all possible classification thresholds even though many of these are unacceptable and would never be considered in practice [37].

20.6 What Are the Next Steps?

The SICULA should be compared to more recent severity scores. Nonetheless, such scores (e.g., SAPS 3 and APACHE III) have been reported to face the same drawbacks as SAPS II [9, 12, 38]. Moreover, those scores remain the most widely used scores in practice [39]. Despite the fact that MIMIC II encompasses data from multiple ICUs, the sample still comes from a single hospital and thus needs further external validation. However, the patients included in the MIMIC-II cohort seem representative of the overall ICU patient population, as reflected by a hospital mortality rate in the MIMIC-II cohort that is similar to the one reported for ICU patients during the same time period [40]. Consequently, our score can be reasonably expected to exhibit, in other samples, performance characteristics similar to those reported here, at least in samples drawn from similar patient populations. A large representation in our sample of CCU or CSRU patients, who often have lower severity scores than medical or surgical ICU patients, may have limited our score's applicability to more critically ill patients. Finally, a key assumption justifying this study was that the poor calibration associated with current 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. Expanding the set of explanatory variables used could potentially result in a score with even better predictive performance. In the future, expending the number of explanatory variables will probably further improve the predictive performances of the score.

20.7 Conclusions

Thanks to a large collection of potential predictors and a sufficient sample size, MIMIC II dataset offers a unique opportunity to develop and validate new severity scores. In this population, the prediction of hospital mortality based on the Super Learner achieves significantly improved performance, both in terms of calibration and discrimination, as compared to conventional severity scores. The SICULA prediction algorithm is a promising alternative that could prove valuable in clinical practice and for research purposes. Externally validating results of this study in different populations (especially population outside the U.S.), providing regular

update of the SICULA fit and assessing the potential benefit of including additional variables in the score remain important future challenges that are to be faced in the second stage of the SICULA project.

Code Appendix

This case study used code from the Super Learner Library, implemented in R. Further details and code are available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-data-book. The following algorithms are included in the Super Learner Library.

Parametric algorithms:

- Logistic regression: standard logistic regression, including only main terms for each covariate and including interaction terms [42] (SL.glm),
- Stepwise regression: logistic regression using a variable selection procedure based on the Akaike Information Criteria [43] (SL.stepAIC),
- Generalized additive model [43] (SL.gam):,
- Generalized linear model with penalized maximum likelihood [44] (SL.glmnet),
- Multivariate adaptive polynomial spline regression [44] (SL.polymars),
- Bayesian generalized linear model [45] (SL.bayesglm).

Non parametric algorithms:

- Random Forest [46] (SL.randomForest),
- Neural Networks [47] (SL.nnet),
- Bagging classification trees [48] (SL.ipredbagg),
- Generalized boosted regression model [49] (SL.gbm),
- Pruned Recursive Partitioning and Regression Trees [50] (SL.rpartPrune),
- Bayesian Additive Regression Trees [51] (SL.bart).

References

1. Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE (1981) APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 9(8):591–597
2. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: a severity of disease classification system. Crit Care Med 13(10):818–829
3. Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, Mercier P, Thomas R, Villers D (1984) A simplified acute physiology score for ICU patients. Crit Care Med 12 (11):975–977
4. Le Gall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 270(24):2957–2963
5. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J (1993) Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. JAMA 270(20):2478–2486
6. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A (1991) The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 100(6):1619–1636
7. Le Gall JR, Neumann A, Hemery F, Bleriot JP, Fulgencio JP, Garrigues B, Gouzes C, Lepage E, Moine P, Villers D (2005) Mortality prediction using SAPS II: an update for French intensive care units. Crit Care 9(6):R645–R652
8. Nassar AP, Jr, Mocelin AO, Nunes ALB, Giannini FP, Brauer L, Andrade FM, Dias CA (2012) Caution when using prognostic models: a prospective comparison of 3 recent prognostic models. J Crit Care 27(4), 423.e1–423.e7
9. Poole D, Rossi C, Latronico N, Rossi G, Finazzi S, Bertolini G (2012) 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 38(8):1280–1288
10. Metnitz B, Schaden E, Moreno R, Le Gall J-R, Bauer P, Metnitz PGH (2009) Austrian validation and customization of the SAPS 3 admission score. Intensive Care Med 35(4):616–622
11. Moreno RP, Metnitz PGH, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall J-R (2005) 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 31(10):1345–1355
12. Beck DH, Smith GB, Pappachan JV, Millar B (2003) External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. Intensive Care Med 29(2):249–256
13. Aegerter P, Boumendil A, Retbi A, Minvielle E, Dervaux B, Guidet B (2005) SAPS II revisited. Intensive Care Med 31(3):416–423
14. Ledoux D, Canivet J-L, Preiser J-C, Lefrancq J, Damas P (2008) SAPS 3 admission score: an external validation in a general intensive care population. Intensive Care Med 34(10):1873–1877
15. Dybowski R, Weller P, Chang R, Gant V (1996) Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet 347(9009):1146–1150
16. Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT (2001) Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. Crit Care Med 29(2):291–296
17. Ribas VJ, López JC, Ruiz-Sanmartin A, Ruiz-Rodríguez JC, Rello J, Wojdel A, Vellido A (2011) Severe sepsis mortality prediction with relevance vector machines. Conf Proc IEEE Eng Med Biol Soc 2011:100–103
18. Kim S, Kim W, Park RW (2011) A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Health Inform Res 17(4):232–243

19. Foltran F, Berchialla P, Giunta F, Malacarne P, Merletti F, Gregori D (2010) Using VLAD scores to have a look insight ICU performance: towards a modelling of the errors. J Eval Clin Pract 16(5):968–975
20. Gortzis LG, Sakellaropoulos F, Ilias I, Stamoulis K, Dimopoulou I (2008) Predicting ICU survival: a meta-level approach. BMC Health Serv Res 8:157–164
21. Dudoit S, Van Der Laan MJ (2003) Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. Statistical Methodology 2(2):131–154
22. Lee J, Scott DJ, Villarroel M, Clifford GD, Saeed M, Mark RG (2011) Open-access MIMIC-II database for intensive care research. Conf Proc IEEE Eng Med Biol Soc 2011:8315–8318
23. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 39(5):952–960
24. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101(23): E215–E220
25. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ (2015) Mortality prediction in intensive care units with the super ICU learner algorithm (SICULA): a population-based study. Lancet Respir Med 3(1)
26. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart CK, Suter PM, Thijs LG (1996) 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 22 (7):707–710
27. Van Der Laan MJ, Dudoit S (2003) Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper, no 130, pp 1–103
28. van der Laan MJ, Polley EC, Hubbard AE (2007) Super learner. Stat Appl Genet Mol Biol 6:25
29. Cox DR (1958) Two further applications of a model for binary regression. Biometrika 45 (3/4):562–565
30. Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K (2006) 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 34(5):1378–1388
31. Kramer AA, Zimmerman JE (2007) Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. Crit Care Med 35(9):2052–2056
32. Bertolini G, D'Amico R, Nardi D, Tinazzi A, Apolone G (2000) One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. J Epidemiol Biostat 5(4):251–253
33. Miller ME, Hui SL, Tierney WM (1991) Validation techniques for logistic regression models. Stat Med 10(8):1213–1226
34. Cook NR (2007) Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 115(7):928–935
35. Cook NR (2008) Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem 54(1):17–23
36. Pencina MJ, D'Agostino RB, Sr, D'Agostino RB, Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 27(2):157–172; discussion 207–212, Jan 2008
37. Greenland S (2008) 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 10.1002/sim.2929. Stat Med 27(2):199–206

38. Sakr Y, Krauss C, Amaral ACKB, Réa-Neto A, Specht M, Reinhart K, Marx G (2008) 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 101(6):798–803
39. Rosenberg AL (2002) Recent innovations in intensive care unit risk-prediction models. Curr Opin Crit Care 8(4):321–330
40. Zimmerman JE, Kramer AA, Knaus WA (2013) Changes in hospital mortality for United States intensive care unit admissions from 1988 to 2012. Crit Care 17(2):R81
41. Van der Laan MJ, Rose S (2011) Targeted learning: causal inference for observational and experimental data. Springer, Berlin
42. McCullagh P, Nelder JA (1989) Generalized linear models, vol 37. Chapman & Hall/CRC
43. Venables WN, Ripley BD (2002) Modern applied statistics with S. Springer, Berlin
44. Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 1–67
45. Gelman A, Jakulin A, Pittau MG, Su YS (2008) A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat 1360–1383
46. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
47. Ripley BD (2008) Pattern recognition and neural networks. Cambridge university press, Cambridge
48. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
49. Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77(4):802–813
50. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Chapman & Hall, New York
51. Chipman HA, George EI, McCulloch RE (2010) BART: Bayesian additive regression trees. Ann Appl Stat 4(1):266–298