Chapter 21
Mortality Prediction in the ICU

Joon Lee, Joel A. Dubin and David M. Maslove

Learning Objectives
Build and evaluate mortality prediction models.

1. Learn how to extract predictor variables from MIMIC-II.
2. Learn how to build logistic regression, support vector machine, and decision tree models for mortality prediction.
3. Learn how to utilize adaptive boosting to improve the predictive performance of a weak learner.
4. Learn how to train and evaluate predictive models using cross-validation.

21.1 Introduction

Patients admitted to the ICU suffer from critical illness or injury and are at high risk of dying. ICU mortality rates differ widely depending on the underlying disease process, with death rates as low as 1 in 20 for patients admitted following elective surgery, and as high as 1 in 4 for patients with respiratory diseases [1]. The risk of death can be approximated by evaluating the severity of a patient's illness as determined by important physiologic, clinical, and demographic determinants.

In clinical practice, estimates of mortality risk can be useful in triage and resource allocation, in determining appropriate levels of care, and even in discussions with patients and their families around expected outcomes. Estimates of mortality risk are, however, based on studying aggregate data from large, heterogeneous groups of patients, and as such their validity in the context of any single patient encounter cannot be assured. This shortcoming can be mitigated by

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_21

personalized mortality risk estimation, which is well discussed in [2, 3], but is not a subject of the present study.

Perhaps even more noteworthy uses of mortality prediction in the ICU are in the areas of health research and administration, which often involve looking at cohorts of critically ill patients. Traditionally, such population-level studies have been more widely accepted as applications of mortality prediction given the cohort-based derivation of prediction models. In this context, mortality prediction is used to compare the average severity of illness between groups of critically ill patients (for example, between patients in different ICUs, hospitals, or health care systems) and between groups of patients enrolled in clinical trials. Predicted mortality can be compared with observed mortality rates for the purpose of benchmarking and performance evaluation of ICUs and health systems.

A number of severity of illness (SOI) scores have been introduced in the ICU to predict outcomes including death. These include the APACHE scores [4], the Simplified Acute Physiology Score (SAPS) [5], the Mortality Probability Model (MPM) [6], and the Sequential Organ Failure Assessment (SOFA) score [7]. These scoring systems perform well, with areas under the receiver operator characteristic (ROC) curves (AUROCs) typically between 0.8 and 0.9 [5, 6, 8]. Current research is exploring ways to leverage the enhanced completeness and expressivity of modern electronic medical records (EMRs) in order to improve prediction accuracy. In particular, the granular nature (i.e., a rich set of clinical variables recorded in high temporal resolution) of EMRs can lead to creating a personalized predictive model for a given patient by identifying and utilizing data from similar patients.

21.2 Study Dataset

This case study aimed to create mortality prediction models using the first ICU admissions from all adult patients in MIMIC-II version 2.6. In the *icustay_detail* table, adult patients in MIMIC-II can be identified by *icustay_age_group='adult'*, whereas the first ICU admission of each patient can be selected by *subject_icustay_seq=1*. In addition, all ICU stays with a null *icustay_id* were excluded, since *icustay_id* was used to find the data in other tables that correspond to the included ICU stays. A total of 24,581 ICU admissions in MIMIC-II met these inclusion criteria.

The following demographic/administrative variables were extracted to be used as predictors: age at ICU admission, gender, admission type (elective, urgent, emergency), and first ICU service type of the ICU admission. Furthermore, the first measurement in the ICU of the following vital signs and lab tests was each extracted as a predictor: heart rate, mean and systolic blood pressure (invasive and noninvasive measurements combined), body temperature, SpO_2, respiratory rate, creatinine, potassium, sodium, chloride, bicarbonate, hematocrit, white blood cell count, glucose, magnesium, calcium, phosphorus, and lactate. Although the very

first measurements in the ICU were extracted, the exact measurement time with respect to the ICU admission time would have varied between patients. Also, this approach to variable-by-variable data extraction does not ensure concurrent measurements within patient. For the vast majority of the ICU admissions in MIMIC-II, however, measurements of these common clinical variables were obtained at the beginning of the ICU admission, or at most within the first 24 h.

As the patient outcome to be predicted, mortality at 30 days post-discharge from the hospital was extracted. In MIMIC-II, this binary outcome variable can be obtained by comparing the date of death (found in the *d_patients* table) and the hospital discharge date (found in the *icustay_detail* table). If our focus were on a greater time period to post-discharge death, we would have extracted mortality date in an attempt to predict survival time.

21.3 Pre-processing

Some of the extracted variables require further processing before they can be used for predictive modeling. In MIMIC-II, some ages are unrealistically large ($\sim$200 years), as they were intentionally inserted to mask the actual ages of those patients who were 90 years or older and still alive (according to the latest social security death index data), which is protected health information. For these patients, the median of such masked ages (namely, 91.4) was substituted. Furthermore, regarding ICU service type, FICU (Finard ICU; this is a term specific to Beth Israel Deaconess Medical Center where MIMIC-II data were collected) was converted to MICU (medical ICU) since there are only a small number of FICU admissions in MIMIC-II and FICU is nothing more than a special MICU.

There are abundant missing data in MIMIC-II. Although there are ways to make use of ICU admissions with incomplete data (e.g., imputation), this case study simply excluded cases with incomplete data since missing data is discussed in depth in [insert reference to Missing Data Chapter, Part 2]. After exclusion of cases with incomplete data, only 9269 ICU admissions remained. This still is a sufficient sample size to conduct the present case study, but approaches such as imputation and/or exclusion of variables with frequent missing data should be considered if a larger patient sample size is required.

With default settings in R, numeric variables are normally imported correctly with proper handling of missing data (flagged as NA), but special care may be needed for importing categorical variables. In order to avoid the empty field being imported as a category on its own, this case study (1) imported the categorical variables as strings, (2) converted all empty fields to NA, and then (3) converted the categorical variables to factors. This case study includes the following categorical variables: gender, admission type, ICU service type, and 30-day mortality.

21.4 Methods

The following predictive models were employed: logistic regression (LR), support vector machine (SVM), and decision tree (DT). These models were chosen due to their widespread use in machine learning. Although the reader should refer to appropriate chapters in Part 2 to learn more about these models, a brief description of each model is provided here.

LR is a model that can learn the mathematical relationship, within a restricted framework using a logistic function, between a set of covariates (i.e., predictor variables in this case study) and a binary outcome variable (i.e., mortality in this case study). Once this relationship is learned, the model can make a prediction for a new case given the predictor values from the new case. LR is very widely used in health research thanks to its easy interpretability.

SVMs are similar to LR in the sense that it can classify (or predict) a given case in terms of the outcome, but they do so by coming up with an optimal decision boundary in the data space where the dimensions are the covariates and all available data points are plotted. In other words, SVMs attempt to draw a decision boundary that puts as many negative (survived) cases as possible on one side of the boundary and as many positive (expired) cases as possible on the other side.

Lastly, DTs have a tree-like structure that consists of decision nodes in a hierarchy. Each decision node leads to two branches depending on the value of a particular covariate (e.g., age >65 or not). Each case follows appropriate branches until it reaches a terminal leaf node which is associated with a particular outcome. DT learning algorithms automatically learn an optimal decision tree structure given a set of data.

We also attempted to improve the predictive performance of the DT by applying adaptive boosting, i.e., AdaBoost [9]. AdaBoost can effectively improve a weak predictive model by building an ensemble of models that progressively focus more on the cases that are inaccurately predicted by the previous model. In other words, AdaBoost allowed us to build a series of DTs where the ones built later were experts on more challenging cases. In AdaBoost, the final prediction is the average of the predictions from the individual models.

In order to run the provided R code, the following R packages should be installed via *install.packages()*: *e1071*, *ada*, *rpart*, and *ROCR*. The training functions for LR, SVM, and DT are *glm()*, *svm()*, and *rpart()*, respectively. For all models, default parameter settings were used.

For training and testing, 10-fold cross-validation was utilized. Under such a scheme, the ICU admissions included in the case study were randomly partitioned into 10 similarly sized groups (a.k.a. folds). The procedure rotated through the 10 folds to train predictive models based on 9 folds (training data) and test them on the remaining fold (test data), until each fold is utilized as test data.

Predictive performance was measured using AUROC which is a widely used performance metric for binary classification. For each predictive model, the

AUROC was calculated for each fold of the cross-validation. In the provided R code, the *comp.auc()* function is called to calculate the AUROC given a set of predicted probabilities from a model and the corresponding actual mortality data.

21.5 Analysis

The following were the AUROCs of the predictive models (shown in mean [standard deviation]): LR—0.790 [0.015]; SVM—0.782 [0.014]; DT—0.616 [0.049]; AdaBoost—0.801 [0.013]. Hence, in terms of mean AUROC, AdaBoost resulted in the best performance, while DT was clearly the worst predictive model. DT was only moderately better than random guessing (which would correspond to an AUROC of 0.5) and as a result can be considered a weak learner. Note that AdaBoost was able to substantially improve DT, which is consistent with its known ability to effectively improve weak learners. Because of the random data partitioning of cross-validation, slightly different results will be produced every time the provided R code is run. Using *set.seed()* in R can seed the random number generation in *sample()* and make the results reproducible, but this was not used in this case study for a more robust evaluation of the results.

As a comparison, a previous study [2] reported mean AUROCs of 0.658 (95 % confidence interval (CI): [0.648,0.668]) and 0.633 (95 % CI: [0.624,0.642]) for SAPS I and SOFA, respectively, for predicting 30-day mortality for 17,152 adult ICU stays in MIMIC-II, despite that the analyzed patient cohort was a bit different from the one in this case study. More advanced SOI scores such as APACHE IV would have achieved a comparable or better performance than the predictive models investigated in this case study (only SAPS I and SOFA are available in MIMIC-II), but it should be noted that those advanced SOI scores tend to use a much more comprehensive set of predictors than the ones used in this case study.

21.6 Visualization

Figure 21.1 shows the performances of the predictive models in a boxplot. It is visually apparent that AdaBoost, LR, and SVM resulted in similar performance, while DT yielded not only the worst performance but also the largest variability in AUROC, which sheds light on its sensitivity to the random data partitioning in cross-validation.

Figure 21.2 is an interesting visualization of the prediction results, where each circle represents a patient and the color of the circle indicates the prediction result (correct or incorrect) of the patient. Random horizontal jitter was added to each point (this simply means that a small random shift was applied to the x-value of each point) to reduce overlap with other points. Prediction results from only one of the ten cross-validation folds are shown, with a threshold of 0.5 (arbitrarily selected;

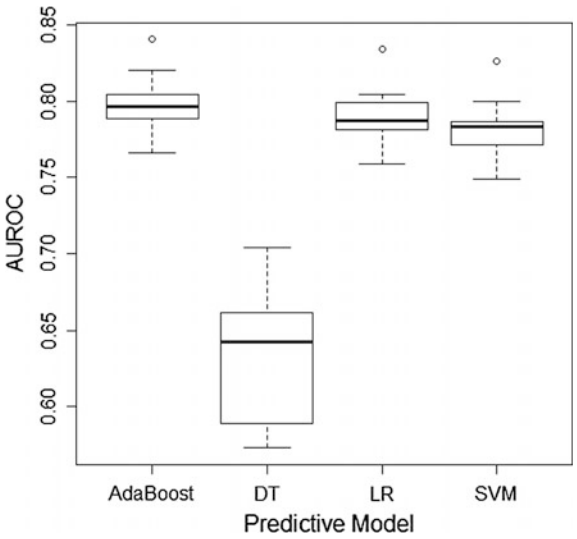

Fig. 21.1 A *box* and *whisker* plot showing mortality prediction performances of several predictive models from 10-fold cross-validation. *AUROC* Area under the receiver operating characteristic curve; *DT* Decision tree; *LR* Logistic regression; *SVM* Support vector machine

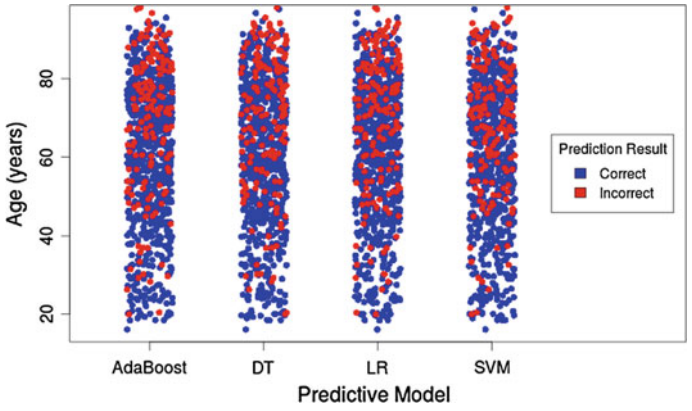

Fig. 21.2 Prediction results for individual patients as a function of age, stratified by predictive model. Results from only one of the ten cross-validation folds are plotted here

the reader may be interested in studying how this threshold affects this figure) applied to the estimated mortality risks from the predictive models (by calling the *th.pred()* function in the R code). Figure 21.2 shows the prediction results as a function of age, but the variable on the y-axis can easily be changed to some other variable of interest (e.g., heart rate, creatinine). One observation that is clear in Fig. 21.2 but not in Fig. 21.1 is that predictive accuracy is higher for younger

patients (e.g., <40 years) than for older patients, across all predictive models. This is most likely due to the fact that mortality rate is much lower among younger patients than older patients, and predictive models can achieve a high accuracy by biasing towards predicting low mortality risks (however, this would lead to a low sensitivity). Hence, it is important to note that although Fig. 21.2 conveys a sense of overall accuracy, it does not reveal sensitivity, specificity, positive predictive value, or negative predictive value.

21.7 Conclusions

Using clinical and demographic data from the MIMIC II database, this case study used machine learning algorithms to classify patients as alive or dead at 30 days after hospital discharge. Results were comparable to those obtained by the most up to date SOI scores currently in use. Unlike these scores, however, the learning algorithms used did not have access to specific diagnoses and procedures, which can add considerable predictive power. An advantage of using only clinical and demographic data, however, is that they are more routinely available and as a result predictive models based on them can be used more widely. Moreover, our algorithms were applied to an undifferentiated population of critically ill patients, rather than tailored to specific groups such as those following cardiovascular surgery (i.e., cardiac surgery recovery unit (CSRU) patients), which has also been shown to enhance predictive performance [3]. The success of prediction seen in this case study likely reflects the power of the learning algorithms used, as well as the utility of both the size and granularity of the database studied.

One useful prospect that leverages the dynamic nature of EMR data is the potential to update training data and prediction models as the most recent clinical data become available. This would theoretically lead to equally dynamic scoring systems that generate more accurate predictions by reflecting current practices. A trade-off becomes apparent between the use of the most current data, which is likely to be the most representative, and the inclusion of older data as well, which may be less relevant but provides greater statistical power.

21.8 Next Steps

Although AUROCs near 0.8 represent good performance, the fact that LR, SVM, and AdaBoost resulted in similar performance may imply that performance could be limited by the predictor variables rather than model selection. A meaningful future study could further investigate predictor selection or different representations of the same variables (e.g., temporal patterns rather than measurements at a specific time point; see the Hyperparameter Selection chapter of Part 3).

Since the default parameter settings were used for the LR, SVM, DT, and AdaBoost, another reasonable next step is to investigate how changing the parameters affect predictive performance. Please refer to R Help or appropriate R package documentation to learn more about the model parameters.

To improve predictive performance, we have previously considered a personalized mortality prediction approach where only the data from patients that are similar to an index patient (for whom prediction is to be made) are used for training customized predictive models [2]. Using a particular cosine-similarity-based patient similarity metric and LR, the maximum AUROC this study reported was 0.83. In light of this promising result, the reader is invited to pursue similar personalized approaches with new patient similarity metrics.

Bayesian methods [10] offer another prediction paradigm that may be worth investigating. Bayesian methods strike a balance between subject-matter expertise (for mortality prediction in the ICU, this would correspond to clinical expertise regarding mortality risk) and empirical evidence in the clinical data. Since the machine learning models discussed in this chapter were purely empirical, the explicit addition of clinical expertise through the Bayesian paradigm can potentially improve predictive performance.

Aside from AUROC, there are other ways to evaluate predictive performance, including the scaled Brier score. Please see [11] for more information. Once a threshold is applied to predicted mortality risk, more conventional performance measures such as accuracy, sensitivity, specificity, etc. can also be calculated. Since each performance measure has pros and cons (e.g., while AUROC provides a more complete assessment than simple accuracy, it becomes biased for skewed datasets [12]), it may be best to calculate a variety of measures for a holistic assessment of predictive performance.

Lastly, data quality is often overlooked but plays an important role in determining what predictive performance is possible with a given set of data. This is a particularly critical issue with retrospective EMR data, the recording of which may have had minimal data quality checks. Implementation of more rigorous data quality checks (e.g., outliers, physiologic feasibility) prior to predictive model training is a meaningful next step.

21.9 Connections

While this chapter focused on mortality prediction, the data extraction and analytic techniques discussed here are widely applicable to prediction of other discrete (e.g., hospital re-admission) and continuous (e.g., length of stay) patient outcomes. In addition, the nuances related to MIMIC-II such as handling ages near 200 years and the service type FICU are important issues for any MIMIC-II study.

The machine learning models (LR, DT, SVM) and techniques (cross-validation, AdaBoost, AUROC) are widely used in a variety of prediction, detection, and data

mining applications, not only in but beyond medicine. Furthermore, given that R is one of the most popular programming languages in data science, being able to manipulate EMR data and apply machine learning in R is an invaluable skill to have.

Code Appendix

The code used in this case study is available from the GitHub repository accompanying this book: https://github.com/MIT-LCP/critical-data-book. Further information on the code is available from this website. The reader can reproduce the present case study by running the following SQL and R codes verbatim:

- `query.sql`: used to extract data from the MIMIC II database.
- `analysis.R`: used to perform data processing.

References

1. Kuzniewicz MW, Vasilevskis EE, Lane R, Dean ML, Trivedi NG, Rennie DJ, Clay T, Kotler PL, Dudley RA (2008) Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. Chest 133(6):1319–1327
2. Lee J, Maslove DM, Dubin JA (2015) Personalized mortality prediction driven by electronic medical data and a patient similarity metric. PLoS ONE 10(5):e0127428
3. Lee J, Maslove DM (2015) Customization of a severity of illness score using local electronic medical record data. J. Intensive Care Med, 0885066615585951
4. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: a severity of disease classification system. Crit Care Med 13(10):818–829
5. Legall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPS-II) based on a european north-american multicenter study. Jama-J Am Med Assoc 270:2957–2963
6. 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

7. Vincent J, Moreno R, Takala J, Willatts S, De Mendonca A, Bruining H, Reinhart C, Suter P, Thijs L (1996) The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Intensive Care Med 22(7):707–710
8. Gursel G, Demirtas S (2006) Value of APACHE II, SOFA and CPIS scores in predicting prognosis in patients with ventilator-associated pneumonia. Respiration. 73(4):503–508
9. Freund Y, Schapire R (1995) A desicion-theoretic generalization of on-line learning and an application to boosting. Comput Learn Theory 55(1):119–139
10. Gelman A, Carlin JB, Stern HS, Rubin DB (2014) Bayesian data analysis, vol 2. Taylor & Francis, UK
11. Wu YC, Lee WC (2014) Alternative performance measures for prediction models. PLoS One 9(3)
12. Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning—ICML'06, pp 233–240