Predicting In-Hospital Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012

Ikaro Silva[1], George Moody[1], Daniel J Scott[1], Leo A Celi[1,2], Roger G Mark[1,2]

[1] Massachusetts Institute of Technology, Cambridge, USA
[2] Beth Israel Deaconess Medical Center, Boston, USA

Abstract

Acuity scores, such as APACHE, SAPS, MPM, and SOFA, are widely used to account for population differences in studies aiming to compare how medications, care guidelines, surgery, and other interventions impact mortality in Intensive Care Unit (ICU) patients. By contrast, the focus of the PhysioNet/CinC Challenge 2012 is to develop methods for patient-specific prediction of in-hospital mortality. The data used for the challenge consisted of 5 general descriptors and 36 time series (measurements of vital signs and laboratory results) from the first 48 hours of the first available ICU stay of 12,000 adult patients from the MIMIC II database. The challenge was organized as two events: event 1 measured performance of a binary classifier, and event 2 measured performance of a risk estimator. The score of event 1 was the lower of sensitivity and positive predictive value. The score for event 2 was a range-normalized Hosmer-Lemeshow statistic. A baseline algorithm (using SAPS-1) obtained event 1 and 2 scores of 0.3125 and 68.58 respectively. Most participants submitted entries that outperformed the baseline algorithm. The top final scores for events 1 and 2 were 0.5353 and 17.88 respectively.

1. Introduction

The development of methods for prediction of mortality rates in Intensive Care Unit (ICU) populations has been motivated primarily by the need to compare the efficacy of medications, care guidelines, surgery, and other interventions when, as is common, it is necessary to control for differences in severity of illness or trauma, age, and other factors. For example, comparing overall mortality rates between trauma units in a community hospital, a teaching hospital, and a military field hospital is likely to reflect the differences in the patient populations more than any differences in standards of care. The use of acuity scores such as SAPS aims to compensate for population differences in order to compare practice variations objectively. This challenge, however, sought to encourage development of methods for *patient-specific* prediction of in-hospital mortality, making use of not only the parameters used to compute SAPS scores, but also other observations including time series of vital signs during the 48 hours following ICU admission. Our hypothesis is that this additional information, and particularly observations of dynamic changes in vital signs (as opposed to a single maximum deranged value), may aid in early identification of patients with elevated risk as well as those whose status may be stable or improving.

2. ICU data

The ICU data used for the challenge were extracted from the MIMIC II Clinical Database, version 2.6 [1]. We selected 12,000 subjects at random from the 12,753 subjects whose age at ICU admission was 16 years or over, and whose initial ICU stay was at least 48 hours long. No other exclusion criteria were applied. We divided these patients randomly into three groups of 4000 (training set A, open test set B, and hidden test set C). For each of these 12,000 patients, we extracted the general descriptors and all observations of the time series variables listed in Table 1 from the first 48 hours of the first ICU stay.

PhysioNet provides free access to the Table 1 data for sets A and B, and the Table 2 (outcome) data for set A only. The remaining Challenge data (set C) have been withheld and were used only to evaluate participants' final algorithms for mortality prediction and risk assessment.

2.1. Input variables

Up to 41 variables were recorded at least once during the first 48 hours after admission to the ICU. Not all variables were available in all cases. Five of these variables were general descriptors collected on admission: age, gender, height, ICU type, and initial weight. The average (standard deviation) for age, uncorrected height, and uncorrected initial weights were 64.5 (17.1) years, 169.5 (17.1) centimeters, and 81.2 (23.8) kg; 43.8% were females, and 56.1% males. The largest number of patients was admitted to the medical ICU (35.8%), followed by the surgical (28.4%), cardiac surgery recovery (21.1%), and coronary (21.1%) ICUs.

The remaining 36 variables were time series (Table 2.1), for which multiple observations could be available. Each observation had an associated time-stamp indicating the elapsed time of the observation since ICU admission in hours and minutes.

2.2. Outcome-related descriptors

Five outcome-related descriptors for the data set A were made available for challenge participants. The descriptors were: SAPS-1 score [2], SOFA score [3], length of stay in days (LOS), length of survival following ICU admission in days (up to 2 years), and in-hospital death (0 = survivor, 1 = died in the hospital). The original SAPS-1 was designed to be calculated on data collected during the first day in the ICU, but because the data set contained incomplete or missing data, the SAPS-1 for the challenge was based on the first period in which all of the SAPS-1 variables were first available (which may include measurements from the second day). A sample challenge entry based on the SAPS-I was provided as an example entry to competitors in both MATLAB and C code. The mean (standard deviation) for all five outcome descriptors on the entire data set (12,000 ICU stays) were: 14.9 (5.2) SAPS-1, 6.4 (4.2) SOFA, 13.4 (12.8) LOS, and 133.9 (372.7) for survival days. The overall mortality rate was 14.2%.

3. Scoring criteria

Due to its unambiguous definition and use in previous similar studies [4–7], we used in-hospital death as the outcome variable to be predicted in the challenge. We defined two challenge events:

Event 1 required participants' algorithms to classify each case as a survivor (at least until discharge from the hospital) or as a non-survivor. The final event 1 score earned by each algorithm was dependent on the counts of true positives (TP), false negatives (FN), and false positives (FP) (Table 2) when tested on set C. We defined sensitivity and positive predictivity as usual:

$$Se = TP/(TP + FN) \tag{1}$$

$$P^{+} = TP/(TP + FP) \tag{2}$$

and defined the event 1 score as the *smaller* of these measures:

$$Score1 = min(Se, P^{+}) \tag{3}$$

This criterion was chosen as a reasonable trade-off between accuracy of discrimination and prognostic value

Table 1. Time-series variables for the challenge and percentage of patients for whom at least one measurement was available during the first 48 ICU hours (total of 12,000 ICU stays).

Measurement	%	Physical Units
ABP (Arterial blood pressure)		
Invasive (diastolic, mean, systolic)	98.4	mmHg
Non-invasive (diastolic)	87.3	mmHg
Non-invasive (mean)	87.2	mmHg
Non-invasive (systolic)	87.6	mmHg
Albumin	40.5	g/dL
ALP (Alkaline phosphatase)	42.4	IU/L
ALT (Alkaline transaminase)	43.4	IU/L
AST (Aspartate transaminase)	43.4	IU/L
Bilirubin	43.4	mg/dL
BUN (Blood urea nitrogren)	98.4	mg/dL
Cholesterol	7.9	mg/dL
Creatinine	98.4	mg/dL
FiO2 (Fractional inspired oxygen)	67.6	[0-1]
Glasgow Coma Score (GCS)	98.4	[3-15]
Glucose	97.5	mg/dL
HCO3 (Serum bicarbonate)	98.2	mmol/L
HCT (Hematocrit)	98.4	%
Heart rate	98.4	bpm
K (Serum potassium)	97.9	mEq/L
Lactate	54.8	mmol/dL
Mg (Serum magnesium)	97.5	mmol/L
Mechanical ventilation	63.1	[yes/no]
Na (Serum sodium)	98.2	mEq/L
PaCO2	75.4	mmHg
PaO2	75.4	mmHg
pH	75.9	[0-14]
Platelets	98.3	cells/nL
Respiration rate	27.7	bpm
SaO2	44.7	%
Temperature	98.4	Celsius
Troponin-I	4.7	ug/L
Troponin-T	21.9	ug/L
Urine output	97.4	mL
WBC (White blood cell count)	98.2	cells/nL
Weight	67.7	kg

Table 2. Definition of discrimination variables used for event 1.

	Observed	
	Deaths	Survivals
Predicted Deaths	TP	FP
Predicted Survivals	FN	TN

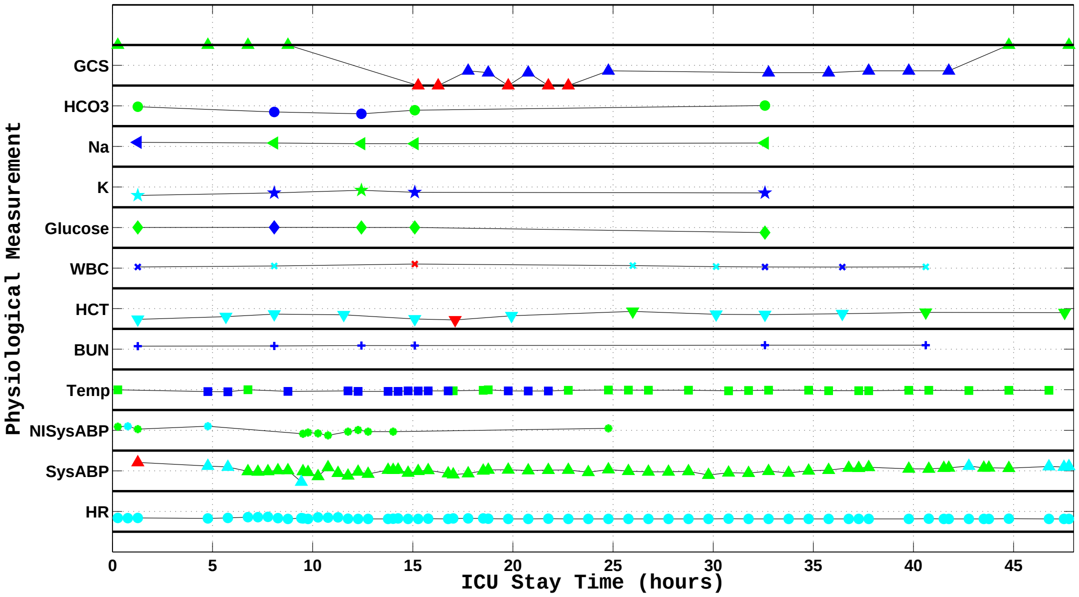

Figure 1. An example of ICU stay data used for the challenge. This particular patient had a SAPS-I = 32, yet survived his hospital stay (see text for details).

given the large skew in class distribution (around 14% deceased vs 86% survivors). Moreover, this choice stimulated competitors to optimize the Precision-Recall (PR) curve of their classifiers, instead of the usual receiver operating curve (ROC). The optimization of the PR curve area has been suggested as a good criterion for unevenly distributed classification problems [8]. Thus the goal for event 1 was to maximize $Score1$.

Event 2 required algorithms to assign an estimate of in-hospital mortality risk to each case.

The scoring for event 2 was based on a modified version of the Hosmer-Lemeshow statistic [9], H. The calculation of the H statistic for an entry was done by first sorting its estimated in-hospital mortality risks for the 4000 set C cases and then binning the corresponding records into deciles designated by $g = 1, 2, 3...10$. The H statistic and the $score2$ values were then calculated as:

$$H = \sum_{g=1}^{10} \frac{(O_g - E_g)^2}{N_g \pi_g (1 - \pi_g) + 0.001} \tag{4}$$

$$Score2 = \frac{H}{\pi_{10} - \pi_1} \tag{5}$$

where for each decile g we have: the observed number of in-hospital deaths O_g, the predicted number of deaths E_g, the number of records N_g ($N_g = 400$ for the challenge), and the mean decile estimated risk π_g. The final score for event 2, $score2$, was then calculated by normalizing the H statistic by the mean risk estimates in the top and bottom deciles. This was done in order to ensure that the risk estimates accurately reflected individual patient risks, rather than simply the risk for the entire population of patients (predicting a constant risk for the entire population yields a low but uninformative H value). The goal for event 2 was to minimize $Score2$.

4. Results

Figure 1 shows an example of the first 48 hours of an ICU stay used in this challenge. The physiological waveforms have been shifted and scaled so that their normal values are in the center of their bins and their extreme values are at the edges (for the Glasgow Coma Score, the higher the value, the closer it is to normal). The time series have been coded according to their instantaneous SAPS-I values using the following coding scheme: green (normal) = 0, blue = 1, cyan = 2, magenta = 3, and red = 4. This particular subject received a final SAPS-I score value of 32 (over 98% chance of death according to the sample entry), yet survived his hospital stay.

Clinical information (not available to the competitors) obtained from the MIMIC II database revealed that this subject was a 83 year old man with a pacemaker admitted to the ICU due to gastrointestinal bleeding. The patient was intubated and administered propofol at hour 10,

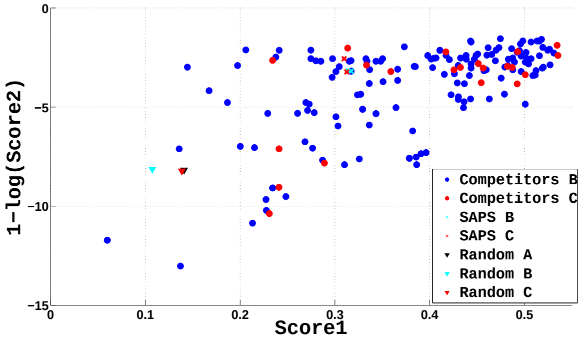

Figure 2. Scores for events 1 and event 2 on set B and C.

received a blood transfusion at around hour 16, and was given warm blanket at around hour 21. He was given antibiotics upon arrival at the ICU and was extubated at around hour 36. Between hours 17 and 25 he was routinely weaned from propofol and brought back to consciousness for cognitive and comfort feedback. This example was chosen in order to highlight some of the difficulties of mortality prediction based on time series analysis. It is possible that other features beyond maximum derangement from normal values may provide prognostic information, such as overall trend and coupling of changes between different measurement variables. Nevertheless, this example also shows that medical conditions (i.e., pacemakers) or interventions can shift the measurements towards a normal or abnormal range, biasing estimates of a patient's genuine state.

Figure 2 shows a scatter plot for the scores of the entries on events 1 and 2. A total of 37 different teams across the world competed in this year's PhysioNet challenge, submitting around 200 different entries for predicting in hospital mortality. Five reference scores were also plotted: guessing with an assumed mortality rate of 14% (triangles), and SAPS-I sample entries (crosses). The top competitors for the challenge achieved significantly better scores with respect to both random guessing and the SAPS-I entries. The approaches used by the competitors went beyond the typical logistic regression used in mortality prediction scores, including support vector machines, neural networks, random forests, and ensemble learning.

5. Discussion

Given that the data sets were created from a diverse population with a wide variety of life-threatening conditions, with frequent missing and occasionally incorrectly recorded observations, idiosyncrasies of care administration, and highly unbalanced class sizes, we expected this Challenge to be difficult. Moreover, as also noted in [7], certain physiological measurements, such as systolic blood pressure, can be more reflective of medical intervention than the genuine state of the patient per se.

The Challenge data sets remain freely available from PhysioNet as a basis for objective comparisons of mortality predictors in future studies.

Acknowledgments

This work was funded in part by the National Institute of Biomedical Imaging and Bio-engineering and by the National Institute of General Medical Sciences, under NIH cooperative agreement U01-EB-008577 and NIH grant R01-EB-001659.

References

[1] Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. Critical Care Medicine 2011;39(5):10.

[2] Le Gall J, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, Mercier P, Thomas R, Villers D. A simplified acute physiology score for ICU patients. Critical Care Medicine 1984;12(11):975–977.

[3] Ferreira F, Bota D, Bross A, Mlot C, Vincent J. Serial evaluation of the SOFA score to predict outcome in critically ill patients. JAMA 2001;286(14):1754–1758.

[4] Le Gall J, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 1993;270(24):2957–2963.

[5] Lemeshow S, Teres D, Klar J, Avrunin J, Gehlbach S, Rapoport J. Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. JAMA 1993;270(24):2478–2486.

[6] Elixhauser A, Steiner C, Harris Robert D, Coffey Rosanna M. Commorbidity measures for use with administrative data. Medical Care 1998;36(1):8–27.

[7] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: A severity of disease classification system. Critical Care 1985;13(10):818–829.

[8] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning ICML 06 2006; 10(2):233–240.

[9] Hosmer D, Lemeshow S. Applied Logistic Regression. Wiley, 2000.

Address for correspondence:

Ikaro Silva
MIT Room E25-505A, Cambridge, MA 02139 USA
ikaro@mit.edu