An Imputation-Enhanced Algorithm for ICU Mortality Prediction

Cheng H Lee, Natalia M Arzeno, Joyce C Ho, Haris Vikalo, Joydeep Ghosh

The University of Texas at Austin, Austin, TX, USA

Abstract

ICU patients are vulnerable to in-ICU morbidities and mortality, making accurate systems for identifying at-risk patients a necessity for improving clinical care. Here, we present an improved model for predicting in-hospital mortality using data collected from the first 48 hours of a patient's ICU stay.

We generated predictive features for each patient using demographic data, the number of observations for each of 37 time-varying variables in hours 0–48 and 47–48 of the stay, and the last observed value for each variable. Missing data are a common problem in clinical data, and we therefore imputed missing values using the mean value for a patient's age and gender group.

After imputing the missing data, we trained a logistic regression using this feature set. We evaluated model performance using the two metrics from the 2012 PhysioNet/CinC Challenge; the first measured model accuracy using the minimum of sensitivity and positive predictive value (Event 1), and the second measured model calibration using the Hosmer-Lemeshow H statistic (Event 2). Our model obtained Event 1 and 2 scores of 0.516 and 14.4 for test set B and 0.482 and 51.7 for test set C, respectively, providing better estimates of in-hospital mortality risk than existing methods such as SAPS-I.

1. Introduction

Intensive care unit (ICU) patients are especially vulnerable to adverse events including in-hospital mortality. Therefore, more accurate methods for identifying high-risk patients are a crucial step for improving clinical care. However, development of such systems is complicated by various factors, including the diversity of data collected over the course of each ICU stay and the frequency of missing data for key predictive variables.

The acute physiology and chronic health evaluation (APACHE) [1], mortality probability model (MPM) [2, 3], and simplified acute physiology score (SAPS) [4–6] are among the most commonly used models for predicting risk of mortality in ICU patients [7,8]. Though widely used and having multiple revisions to accommodate changes in patient populations and advances in hospital care, these scoring systems still have key limitations. The most comprehensive and accurate of these scoring systems, APACHE, is a proprietary tool that requires licensing and is heavily dependent on selecting the correct admission diagnosis [8]. MPM and SAPS examine only a few variables, resulting in easy to use but overly simplistic models that might overlook important physiological measurements. While using only a few key data elements to develop predictive models made sense historically, the availability of detailed electronic medical records and modern machine learning methods has made this rationale obsolete. Most importantly, these models are unable to deal with missing data and assume that unobserved parameters are normal, which can result in underpredicted risk [9].

In this paper, we discuss the development of our entry to the 2012 PhysioNet/Computing in Cardiology Challenge, which aims to develop better methods for predicting in-hospital mortality. We describe the construction of several feature sets using data collected during the first 48 hours of a patient's ICU stay, the use of these feature sets for training a range of predictive models, and the evaluation of these models. We provide results for several of our models in the first two rounds of the competition. Finally, we describe our choice of the model for the final competition round and evaluate our final scores.

2. Methods

2.1. Feature sets

We derived our feature sets using training data from 4,000 ICU patients (Set A) that included demographic information (gender, age, height, and ICU type) and time-series measurements for 37 physiological variables. We cleaned this dataset by removing unknown or missing values, indicated by negative values in the data files. Additionally, we discarded entire records for six patients that were either missing gender information or lacked any time-series measurements.

We constructed one feature set (referred to as the "last available measurement" feature set) using the last recorded measurement for each of the 37 time-series variables or

Table 1. Comparative performance of various imputation methods. A logistic regression model was used to predict patient survival in the training (Set A) data, and performance was evaluated using the two scoring metrics used in the 2012 Physionet/CinC Challenge. The table entries indicate the (mean $\pm$ standard deviation) of event scores obtained through 10-fold cross validation.

Imputation method	Event 1 score	Event 2 score
Population mean	0.467 ± 0.051	11.6 ± 5.4
Age/gender mean	0.467 ± 0.051	10.6 ± 4.3
Age/gender median	0.463 ± 0.052	12.1 ± 5.2
SVD	0.468 ± 0.057	11.9 ± 3.3
PPCA	0.460 ± 0.052	13.1 ± 5.6
kNN	0.451 ± 0.050	10.4 ± 4.3

"not available" if the variable was never observed for a given patient. We also included the number of observations for each variable over hours 0–48 and 47–48 as indicators for the overall health of the patient.

We constructed other feature sets (referred to as the "binned measurements" feature set) by aggregating measurements for each variable into 12- or 24-hour bins and using the summary statistics (minimum, maximum, mean, standard deviation, and number of observations) for each variable-bin pair as model features.

In each of these feature sets, several of the variables, including respiratory rate, troponin I, troponin T, and hemoglobin oxygen saturation ($\mathrm{SaO_2}$), were missing in a overwhelming majority of the patients. Rather than discard these features, we converted them to binary features indicating whether any measurement of these variables occurred within the first 48 hours of a patient's ICU stay. We also created a binary feature to indicate whether a patient was ever placed on mechanical ventilation during the first 48 hours of his or her ICU stay.

2.2. Imputation

Inconsistent recording often affects the availability of measurements in clinical data sets. Within our training data, only a small subset of the variables, such as Glasgow Coma Score (GCS), temperature, and heart rate could be found in $> 98\%$ of the patients. For the majority of the time-varying variables (19 of 37), at least 20% of the patients had missing observations. The prevalence of missing data led to feature sets where no patient had a complete set of features. Thus, we evaluated various imputation methods that use measures of central tendency, matrix factorization, or clustering to estimate missing values in our feature sets. Prior research has shown improvements in predictive model performance on incomplete clinical datasets using central tendency or matrix factorization-based imputation techniques [10].

Six approaches of varying complexity were used to estimate the missing observations. The simplest imputation method replaced missing values with the feature mean estimated from the entire patient population. The second method accounted for fundamental physiological differences between genders and among age groups by imputing values for each patient using the mean feature values for the gender and age decade of that patient. Our third method was also based on patient gender and age but used the median of each feature instead of the mean to account for the skewness or non-Gaussian distribution of the data.

Matrix completion approaches to imputation included singular value decomposition (SVD) and probabilistic principal components analysis (PPCA). In SVD-based imputation [11,12], the patient-by-feature matrix is factorized using SVD, and missing values are then replaced by those estimated in the reduced rank matrix reconstructed using the top few eigenvalues. This process is repeated until the change in consecutive matrices falls below a specified threshold. PPCA-based imputation associates a probabilistic model with the observed data through a Gaussian latent variable model, and missing values are imputed using an iterative, expectation maximization algorithm [13, 14]. Clustering-based methods impute missing values based on the average of the k nearest neighbors [11].

Table 1 illustrates the effect the choice of imputation method has on the Event 1 and Event 2 scores of a logistic regression model trained with the last available measurement feature set. Imputing using the feature mean for a patient's gender and age group has the best overall performance and is computationally cheap to implement. Thus, we used this imputation approach to develop all of our models entered into the 2012 Physionet/CinC Challenge.

2.3. Model development

We estimated the performance of various models developed using several combinations of features and algorithms using 10-fold cross validation. Models were evaluated on their estimated Event 1 (the smaller of sensitivity and positive predictive value to measure model accuracy) and Event 2 (Hosmer-Lemeshow H statistic to measure model calibration) scores.

Those models with high Event 1 scores were trained using the entire Set A data set to estimate model parameters and submitted as our Challenge entries to be scored against a testing dataset (Set B). We dealt with patients missing gender information in set B by marginalizing the risk probabilities predicted by a model over gender.

Our final entry scored against an unseen test set (Set C) was constructed by combining the two best performing models evaluated using Set B. This combined model predicted a class label ("Died" or "Survived") using the

Table 2. Our best performing models when evaluated for accuracy (Event 1/E1) and calibration (Event 2/E2) using Set B testing data. Entries for Set A indicate the (mean $\pm$ standard deviation) of each score observed in 10-fold cross validation. Entry 2 was the best performing Event 1 model, while Entry 1 was the best performing Event 2 model.

	Set A		Set B	
Entry	E1	E2	E1	E2
1	0.467$\pm$0.051	10.6$\pm$4.3	0.497	14.4
2	0.451$\pm$0.049	14.1$\pm$7.5	0.516	30.3
5	0.472$\pm$0.062	12.0$\pm$3.4	0.502	25.7
10	0.459$\pm$0.069	35.0$\pm$6.4	0.511	80.2

model with highest Event 1 scoring and a risk probability using the model with the lowest Event 2 score.

3. Results

Table 2 shows our best scoring entries evaluated against test Set B. Entry 1 was an unregularized logistic regression model that used the last available measurement feature set. Entry 2 was an unregularized logistic regression model that used the last available measurement feature set along with the values of diastolic blood pressure (BP), systolic BP, mean arterial BP, FiO_2, GCS, heart rate, temperature, urine, and weight averaged from hours 0–48 and hours 42–48 as additional features. Entry 5 was an $L1$-regularized logistic regression model that used the 24-hour binned measurements feature set. Entry 10 averaged the risk probabilities from an elastic net model [15, 16] using the 12-hour binned measurements and the Entry 2 model.

All of our models outperformed the Set A SAPS-I Event 1 and Event 2 scores (0.296 and 68.4, respectively) provided by the competition organizers. Existing risk scoring systems like SAPS-I only consider the "worst" observed value in a given time period, and our results demonstrate the increase in predictive power provided by expanding the set of features used by a model.

Surprisingly, our Entry 2 model, which only considers the last acquired and average value of each time-series variable, outperformed Entries 5 and 10, which also included minimum and maximum values in the feature set. This result may be an indication of over-fitting in the Entry 5 and 10 models despite the use of regularizers or of the fact that a smaller set of features is all that is necessary to capture a patient's state.

Although the mortality rates differed significantly across ICU types, neither the inclusion of ICU type as a feature nor training separate models for each type of ICU improved our scores significantly. We also explored using forward and reverse stepwise feature selection, support vector machines (SVMs), random forests, boosting, and ensemble models. However, these algorithms only provided a marginal increase in accuracy (Event 1 score) while being more poorly calibrated (increased Event 2 score); therefore, we did not include these models in any of our Challenge submissions.

Based on the test set B results, we developed a combined model for our final submission that used Entry 2 to predict patient outcome (Event 1) and Entry 1 to generate a risk score for evaluating model calibration (Event 2). When run against testing data in Set C, this model produced an Event 1 score of 0.482 and an Event 2 score of 51.7. The significant increase in the Event 2 score for our model on Set C data indicates that either our model is not as well calibrated as we had hoped or that Set C contains a subpopulation of patients with significantly different characteristics than the population in Sets A and B.

We note that while cross validation provides us with reasonable estimates for a model's Event 1 score, it grossly underestimates the Event 2 score. Although the cross validated Event 2 scores for the training data provides hints about the relative performance of the models, this consistent underestimation led to our decision to select models for submission based solely on Event 1 performance. We suspect that this deficiency arises from each cross validation holdout set being one tenth the size of the Set B and Set C data sets coupled with the fact that the Hosmer-Lemeshow H statistic is extremely sensitive to sample size when deviating from perfect calibration [17].

4. Conclusion

This work presented an improved model for predicting in-hospital mortality risk using 37 time-varying variables. The features of the logistic regression model use simple statistics (the last observed value and mean) from clinical measurements, the number of measurements in the first 48 hours, and the last hour combined with conditional mean imputation derived from age and gender to estimate missing data. Our algorithm outperforms the SAPS-I score with respect to positive predictive value and model calibration.

References

[1] Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. Critical Care Medicine May 2006;34(5):1297–1310.

[2] Lemeshow S, Teres D, Pastides H, Avrunin JS, Steingrub JS. A method for predicting survival and mortality of ICU patients using objectively derived weights. Critical Care Medicine July 1985;13(7):519–525.

[3] 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 November 1993;270(20):2478–2486.

[4] Le Gall JR, 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 November 1984;12(11):975–977.

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

[6] Moreno RP, Metnitz PGH, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall JR, on behalf of 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 Medicine August 2005;31(10):1345–1355.

[7] Keegan MT, Gajic O, Afessa B. Severity of illness scoring systems in the intensive care unit. Critical Care Medicine January 2011;39(1):163–169.

[8] Breslow MJ, Badawi O. Severity Scoring in the Critically Ill: Part 1–Interpretation and Accuracy of Outcome Prediction Scoring Systems. Chest January 2012;141(1):245–252.

[9] Afessa B, Keegan MT, Gajic O, Hubmayr RD, Peters SG. The influence of missing components of the Acute Physiology Score of APACHE III on the measurement of ICU performance. Intensive Care Medicine October 2005; 31(11):1537–1543.

[10] Ho JC, Lee CH, Ghosh J. Imputation-Enhanced Prediction of Septic Shock in ICU Patients. In HI-KDD 2012 : ACM SIGKDD Workshop on Health Informatics. 2012; .

[11] Hastie T, Tibshirani R, Sherlock G, Eisen M, Brown P, Botstein D. Imputing missing data for gene expression arrays. Technical Report 1999;.

[12] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics June 2001;17(6):520–525.

[13] Tipping ME, Bishop CM. Probabilistic Principal Component Analysis. J of the Royal Statistical Society Series B Statistical Methodology January 1999;61(3):611–622.

[14] Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics March 2007; 23(9).

[15] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J of Statistical Software February 2010;33(1):1–22.

[16] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Coxs Proportional Hazards Model via Coordinate Descent. J of Statistical Software 2011;39(5):1–13.

[17] Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. Critical Care Medicine September 2007;35(9):2052–2056.

Address for correspondence:

Joydeep Ghosh
1 University Station C0803
The University of Texas at Austin
Austin, TX 78712, USA
ghosh@ece.utexas.edu