ICU Outcome Predictions using Physiologic Trends in the First Two Days

Mehmet Kayaalp

Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

Abstract

Aims: This study aims to accurately predict patient mortality in the ICU. Given all physiologic measurements in the first 48 hours of the ICU stay, the Bayesian model of the study predicts outcome with a posterior probability.

Methods: This study modeled the outcome as a binary random variable dependent on trends of daily physiologic measures of the patient, where trends were conditionally independent given the outcome. A two-day trend is a sequence of two discrete values, one for each day. Each value (low, medium , high or unmeasured) is a function of the arithmetic mean of that measure on the corresponding day.

Results: The prediction performance of the model was measured as the minimum of sensitivity and positive predictive values. The model yielded a score of 0.39 along with a Hosmer-Lemeshow H statistic of 36, which measures calibration. The perfect scores would be 1.0 and 0, respectively.

Conclusion: The prediction performance of the study was an improvement over the established ICU scoring metric SAPS-I, whose score was 0.32. Calibration of the model outputs was comparable to that of SAPS-I.

1. Introduction

In this study, the author developed a Bayesian model to predict the outcome of ICU patients given their physiologic measurements in the first two days of their ICU stay. The dataset comprised measurements, each with a timestamp relative to the time of ICU admission. For example, line **`00:47,HR,78`** denoted that the heart rate of the patient was measured as 78 beats per minute at 00:47. All patients had data up until the time point of **`48:00`**. Patients who had been discharged from ICU (or died) within the first 48 hours were excluded from the cohort.

The training data consisted of 4000 distinct patients from four different ICUs: Coronary Care Unit, Cardiac Surgery Recovery Unit, Medical ICU, and Surgical ICU. The same was true for the test data. Patients did not change their ICUs during their first two days.

This study was part of a larger challenge, PhysioNet / Computing in Cardiology Challenge 2012. The problems, the datasets, the performance metrics were provided by the organizers [1]. The system performance was measured by the organizers on a validation set that was unavailable to system developers.

2. Background

Predicting ICU outcomes has been studied by various research groups [2-13]. Other studies also used constructive induction methods as this author did to generate new variables based on temporal patterns [3]. Such models may continuously monitor the patient and predict near term mortality (such as in the next 24 hours) before the event occurs. An alert from this type of model may be useful to physicians when such an outcome is unexpected.

Others also studied the problem where the temporal distance between the last data point and the ICU outcome was longer than a day. Kayaalp et al [4] showed that the prediction performance of mortality measured as the area under the ROC curve did not decrease much when the ICU discharge time was in a more distant future than one day.

Time series analysis is usually not suitable for predicting ICU outcomes because the data collection times in the clinical settings are usually not as regular as it is required by this class of methods. To overcome this issue many studies raised the temporal granularity of each measurement to one day and assign to each daily measurement a nominal value such as low, normal, high, or unobserved, summarizing the data points collected throughout that day [2,3,5]. Due to their temporal nature, some ICU scoring metrics such as SOFA may be directly applicable to such models [2,3,5,10].

In this study, the author used a Bayesian conditional independence model. These models are known to be prone to miscalibration [13]. Many investigators have tackled the problem, developing various calibration solutions such as model averaging [7], but an effective and consistently reliable solution remains elusive.

3. Methods

As the first step, I discretized all non-categorical data. The data of every patient started with a set of static data points: age, gender, ICU type, height and weight. I converted height and weight to the corresponding body mass index, which is a simpler and perhaps better indicator of the fitness. The static portion of the data were followed by a stream of data, where each line was composed of a timestamp, measurement type and quantity (e.g., **`00:47,HR,78`**).

In this step, I searched for two boundaries for an optimum split of the range of each measurement type. I chose the optimization metric as *min*(*S*, *PPV*), where *S* stands for sensitivity and *PPV* for positive predictive value. If the two boundaries were distinct, the range was split into three subranges: low, medium and high (L, M, and H); otherwise, into two subranges (L and H). Each set of boundaries was tested as long as there were at least 20 cases per subrange. If the mortality rate within a subrange was higher than the mean mortality rate of the cohort, the algorithm predicted death for all cases in that subrange and survival, otherwise.

Below is a typical line, produced by the discretization algorithm.

HR 59.90 104.15 0.22 [+,-,+] [15/78, 423/3367, 111/492]

The label in the first column indicates the measurement type. In this case, it is heart rate. Values in the second and third columns denote the subrange boundaries. The signs in the square brackets tell that the mortality rate increases in subranges L (i.e., HR < 59.9) and H (i.e., HR > 104.15). Ratios in the last square bracket quantify the mortality rate in frequency counts; e.g., in the M subrange, there were 423 deaths out of 3367 cases, corresponding to a mortality rate lower than usual for this cohort. And finally, the value in the fourth column denotes the score (0.22) that this split yielded, which was the maximum score attained by the algorithm for HR.

The algorithm also checked whether not-measuring a particular physiologic entity was associated with a higher or lower mortality rate. I found that it was the case for body mass index and respiration rate measurements.

BMI 20.55 20.55 0.16 [+,-,U+] [28/141, 233/1964, 293/1895]
RespRate 26.11 26.11 0.16 [-,+,U+] [73/1036,14/65,467/2899]

Here U+ indicates unmeasured cases were positively correlated with mortality. Note that unmeasured BMI means that height and/or weight was not reported for the patient.

In the next step, I computed daily mean values for every measurement type for each patient. I converted those values into their discrete counterparts, L, M, or H. If no measurement was made for a given day, I used the label U. If at least one measurement was made in the first two days, the patient would have had one of the 15 possible two-day trends. Recall that the measurement types were excluded from the model, if the trend was UU, which is not one of the 15 trends considered above.

In the next step, I collected the joint frequency counts of each trend and the outcomes. For example, the joint frequency counts of the outcome co-occurred with trend HH on HR were $f(HR_{HH},d) = 78$ and $f(HR_{HH},s) = 233$, where d and s denote death and survival, respectively. Given HR_{HH}, the odds for mortality was 78:233. The corresponding conditional probability $P(d|HR_{HH}) = 0.25$ was significantly higher than the marginal probability of mortality $P(d) = 0.14$. In this study, I estimated probabilities through Laplace smoothing:

$$P(d|HR_{HH}) = \frac{f(d, HR_{HH}) + 1}{f(d, HR_{HH}) + f(s, HR_{HH}) + 2} \qquad (1)$$

For each patient, our algorithm constructed a new patient-specific Bayesian model by including only those trends and variables that were measured in the first two days at the ICU. All models of this study assumed conditional independence, given outcome.

$$P(d|x_1, \ldots, x_n) = \frac{P(d)\prod_{i=1}^{n} P(x_i|d)}{\sum_{O\in\{d,s\}} P(O)\prod_{i=1}^{n} P(x_i|O)} \qquad (2)$$

In Equation 2, O is the outcome variable taking two distinct values death and survival in the denominator. The number of variables and trends n differs in each model depending on the instantiated variables and trends of that particular patient, which are represented in x_i.

Bayesian conditional independence models (a.k.a. naïve Bayes models) are known to be prone to miscalibration. This is especially true for large n; that is, the larger the n, the more extreme the posteriors get. Given n may vary widely from patient to patient, I had to normalize and calibrate them so that the resulting posterior probabilities could be comparable and patient cases could be correctly rank ordered using their posteriors.

In order to normalize these models, I applied the principle of geometric mean computation. When there were $n + 1$ factors in the numerator (as in Equation 2), I raised each factor to a power of $1/(n + 1)$.

Although this normalization step significantly calibrated the posterior, the results were still far from the desirable accuracy.

In the next step, I rank ordered patient cases by their posterior probabilities and partition them into 10 clusters of 400 cases. I associated each patient case with an

observed probability, which I computed using moving averages. For the k^{th} patient case in this rank order, I took into account the outcomes of all patient cases from $k - 50$ to $k + 50$, calculated their overall mortality rate, and estimated the observed mortality probability p_k of the k^{th} case based on this statistic.

I tested various windows sizes (not just ±50) for a better estimate. As the window size gets larger, the function of observed probabilities gets smoother, but not necessarily more accurate. Note that window size decreases as the p_k nears either extreme. For example, the size of the lower window of the third patient from the top was 50, but there were only two additional patients in the upper window; thus, the total sample size to estimate p_3 was 2 + 1 + 50 = 53.

I plotted the observed probability estimates as a function of posterior probabilities. Using the Microsoft Excel 2010 software package, I fitted trend lines to this plot. Since a single trend line cannot be fitted tightly, I did it piecewise. I split the plot into two or three parts and tried to fit a trend line to each split range. I used the corresponding trend line functions to calibrate the posterior probabilities falling to the same range.

4. Results

The system was tested by the organizers of the Physionet Challenge 2012. The test set size was identical to the training set size. I did not have access to the test set; thus, our analysis on the test results here is diminutive.

The predictive performance was measured on the metric *min*(*S*,*PPV*). Our system received a score of 0.39. This score was higher than our baseline measure 0.31 produced by SAPS-I, a well-established ICU scoring metric.

The Challenge 2012 also had a component where the calibration of prediction probabilities was measured using Hosmer-Lemeshow H statistic. My system's calibration performance was measured as 36, which is very close to SAPS-I's calibration performance score 35 measured by the same program.

Due to the lack of available details, I cannot provide further analysis on the results at this point.

5. Conclusions

The author developed (1) a patient-specific Bayesian approach to predict the ICU outcome of the patient using his/her physiologic trends; (2) a discretization technique to optimize the model input; (3) a calibration technique to produce comparable posterior probabilities from different patient-specific models; and (4) a second calibration technique to produce well-calibrated final posteriors.

Given the nature of the challenge, these techniques had to be developed in a relatively short notice; thus, the author made a number of assumptions in various stages and applied them with little prior testing. For example, in search for optimal boundary lines for the physiologic measures, I used the metric *min*(*S*,*PPV*) without having a chance for validating its efficacy. Although it was the metric for measuring the model performance by the organizers, it might have been suboptimal for this subtask.

I also presumed that trends UU might probably be superfluous, since the ICU physicians did not see any reason to order the corresponding tests and adding superfluous features might have lessened the effects of the important ones. But since I have not validated this assumption, it might well be the case that UU trends, at least for certain variables, might have been strengthening the survival chance.

All these untested assumptions can be scrutinized in the future and they may provide opportunities for improvement of the method of this study. In the spirit of the PhysioNet / Computing in Cardiology challenges, the author think that this study is not an end but a beginning of the search for real answers to these tough problems.

Acknowledgements

This study is supported by the Intramural Research Program at NIH.

References

[1] Predicting mortality of ICU patients: the PhysioNet/Computing in Cardiology Challenge 2012. URL: http://physionet.org/challenge/2012/. Accessed on September 2, 2012.

[2] Kayaalp M, Cooper GF, Clermont G. Predicting ICU mortality: a comparison of stationary and nonstationary temporal models. Proc. AMIA Annual Symposium 2000.

[3] Kayaalp M, Cooper GF, Clermont G. Predicting with variables constructed from temporal sequences. Eighth International Workshop on Artificial Intelligence and Statistics 2001.

[4] Rosenberg AL. Recent innovations in intensive care unit risk-prediction models. Current Opinion in Critical Care 2002, 8:321–30.

[5] Kayaalp M. Learning dynamic Bayesian network structures from data. PhD Dissertation, University of Pittsburgh 2003.

[6] Ohno-Machado L, Resnic FS, Matheny ME. Prognosis in critical care. Annu Rev Biomed Eng 2006;8:567–99.

[7] Silva A, Cortez P, Santos MF, Gomes L, Neves J. Mortality assessment in intensive care units via adverse events using artificial neural networks. Artificial Intelligence in Medicine 2006;36:223-34.

[8] Rao RB, Sandilya S, Niculescu RS, Germond C, Rao H. Clinical and financial outcomes analysis with existing hospital patient records. The Ninth International Conference on Knowledge Discovery and Data Mining, 2003.

[9] Verduijn M, Peek N, Rosseel PMJ, de Jonge E, de Mol BAJM. Prognostic Bayesian networks I: Rationale, learning procedure, and clinical use. Journal of Biomedical Informatics 2007;40:609-18.
[10] Toma T, Abu-Hanna A, Bosman R-J. Discovery and inclusion of SOFA score episodes in mortality prediction. Journal of Biomedical Informatics 2007; 40: 649–60.
[11] Gortzis LG, Sakellaropoulos F, Ilias I, Stamoulis K, Dimopoulou I. Predicting ICU survival: A meta-level approach. BMC Health Services Research 2008;8:157-64.
[12] Hug C. Detecting hazardous intensive care patient episodes using real-time mortality models. PhD Dissertation, Massachussets Institute of Technology, 2009.
[13] Bennett PN. Assessing the calibration of naive Bayes' posterior estimates. Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, 2000. Technical Report CMU-CS-00-155.
[14] Wei W, Visweswaran S, Cooper GF. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. J Am Med Inform Assoc 2011;18:370-5.

Address for correspondence.

Mehmet Kayaalp
Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine, National Institutes of Health
8600 Rockville Pike, Bethesda, MD 20894-3828
mehmet@kayaalp.us