Real-time mortality prediction in the Intensive Care Unit

Alistair E. W. Johnson, DPhil[1], Roger G. Mark, MD PhD[1]
[1]Massachussetts Institute of Technology, Cambridge, MA, USA

Abstract

Real-time prediction of mortality for intensive care unit patients has the potential to provide physicians with a simple and easily interpretable synthesis of patient acuity. Here we extract data from a random time during each patient's ICU stay. We believe this sampling scheme allows for the application of the model(s) across a future patient's entire ICU stay. The AUROC of a Gradient Boosting model was high (AUROC=0.920), even though no information about diagnosis or comorbid burden was utilized. We also compare models using data from the first 24 hours of a patient's stay against published severity of illness scores, and find the Gradient Boosting model greatly outperformed the frequently used Simplified Acute Physiology Score II (AUROC = 0.927 vs. 0.809). We nuance this performance with comparison to the literature, provide our interpretation, and discuss potential avenues for improvement.

Introduction

The intensive care unit (ICU) admits severely ill patients in order to provide radical life saving treatment, such as mechanical ventilation. ICUs frequently have a very high staff to patient ratio in order to facilitate for continuous monitoring of all patients and ensure any deterioration in patient condition is detected and corrected for before it becomes fatal; an approach which has been demonstrated to improve outcomes[1]. As a result, the ICU is a data rich environment. A major effort was placed in utilizing this data to both quantify patient health and predict future outcomes, and one of the most immediately relevant outcomes to the ICU is patient mortality. The APACHE system was first published in[2], and provided predictions for patient mortality based upon data collected in the ICU. While the initial system was based off expert rules, later updates used data driven methods[3]. Other prediction systems have also been developed, including the Acute Physiology Score (APS) III[4], Simplified Acute Physiology Score (SAPS)[5], SAPS II[6], the Sequential Organ Failure Assessment (SOFA) score[7], the Logistic Organ Dysfunction Score (LODS)[8], and the Oxford Acute Severity of Illness Score (OASIS)[9]. For a review of severity illness scores in the ICU, see[10,11]. Note that these models were universally agreed to lack sufficient calibration to be used on the individual level[12], and research goals were shifted to quantify the performance of ICUs and hospitals in aggregate.

With the recent advances in both machine learning and hardware for data archiving, research has begun to return to building better prediction models using more detailed granular data. Past models have been limited by technical and practical considerations, often using summary data from an entire day of a patient's ICU stay which was manually documented by trained personnel. Given many hospitals now have electronic data collection as a part of routine clinical practice, a wealth of data is becoming available for use in predictive modeling. The Medical Information Mart for Intensive Care (MIMIC-III) database by[13] is an example of such an archive. MIMIC-III is a large collection of de-identified electronic medical records for over 40,000 patients admitted to the Beth Israel Deaconess Medical Center in Boston, MA, USA between 2001 and 2012.

Hug et al.[14] investigated the use of real-time prediction models on MIMIC-II, an early version of MIMIC containing patient data up to 2008. They extracted observations for 10,066 patients and used a logistic regression model on all observations, achieving a held-out performance Area Under the Receiver Operator Characteristic curve (AUROC) of 0.885. Mortality prediction was the topic of a PhysioNet Computing in Cardiology Challenge in 2012, specifically the prediction of in-hospital mortality for patients who stayed in the ICU for at least 48 hours[15]. The winning entry utilized a tree based ensemble in a Bayesian framework, with components of the trees being updated using Markov chain Monte Carlo, and achieved an AUROC of 0.860[16]. The runner up included an ensemble of six support vector machine models built on balanced subsets of data each using the same set of all positive outcomes[17].

Lehman et al.[18] applied hierarchical Dirichlet Processes to clinical notes during the first 24 hours and found that the addition of extracted topics to a classifier using only severity of illness improved performance (AUROC = 0.82 versus 0.78 with only SAPS I). Ghassemi et al. extracted topics from the notes and combined these topics with static features from MIMIC in a support vector machine to classify in-hospital mortality (AUROC = 0.840)[19]. The authors

furthered this approach by extracting dynamics associated with topics using a multi-task Gaussian process, improving upon the AUROC of SAPS (AUROC = 0.812 vs 0.702)[20]. Caballero et al.[21] created a latent state model incorporating numeric features, text, and modelled topics achieving an AUROC of 0.866 during the first 24 hours on patients in the MIMIC-II database. Finally, Luo et al.[22] proposed an unsupervised feature extraction model using non-negative matrix factorization, focusing on creating an interpretable model (AUROC=0.848).

The use of mortality prediction models to evaluate ICUs as a whole has found great success, both for identifying useful policies and comparing patient populations. Furthermore, as mentioned, much research has advanced the current state of the art in mortality prediction. In this paper, we propose to address the task of predicting individual patient mortality using a distinct sampling scheme. The aim is building a model which could be applied continuously for a patient, which departs from the common analysis framework of the first 24 hours or distinct daily models. This paper will proceed as follows: first, the patient population is defined and data extraction steps are outlined. Common mortality prediction systems in the literature (commonly called severity of illness scores) are then compared to machine learning approaches using data extracted from the first 24 hours of a patient's stay. The framework for evaluating a real-time mortality prediction system is then established and we evaluate various machine learning models in this context, concluding with a discussion on their efficacy and avenues for improvement.

Data

We initially extracted data for 61,533 distinct ICU stays (all stays available in MIMIC-III v1.4). We subsequently excluded patients who met the following criteria: age less than 16 (MIMIC contains neonatal admissions), an ICU stay shorter than four hours, no data present in the patient flowsheet (likely an administrative error resulting in an incorrect ICU admission), and organ donor accounts. We also excluded patients who had an order for limitation of treatment (not full code) on or before ICU admission. Note that we otherwise retained all ICU stays, including readmissions for patients with multiple visits. The final cohort had 50,488 ICU stays. This cohort was used for both experiments (described later). These stays correspond to adult ICU admissions for surgical, medical, neurological, or coronary critical illness. Table 1 provides a description of the study population.

Demographic	**Mean $\pm$ S.D. or percentage**
Age	63.6 $\pm$ 17.2
Male	56%
Height	169.4 $\pm$ 12.75
Weight	81.0 $\pm$ 24.5
BMI	28.4 $\pm$ 8.05
Emergency admission	85.6%
Service	
Medicine	40.6%
Cardiac medicine	15.2%
Cardiac surgery	10.0%
Surgery	8.09%

Table 1: Demographics of the dataset (N=50,488).

Feature Extraction

For each patient's ICU stay, we extracted data from a fixed window of length W (hours) ending at time $t_{i,w}$ (hours) for each patient i. $t_{i,w}$ was set to some value during a patient's ICU stay, i.e. if $t_{i,adm}$ was the time of the patient's ICU admission and $t_{i,dis}$ was the time of the patient's ICU discharge, then $t_{i,w} \in [t_{i,adm}, t_{i,dis}]$. For the initial evaluation of models against severity scores, $t_{i,w} := 24$ hours and $W := 24$ hours, resulting in data being collected from the first day of a patient's ICU stay. For the second evaluation setting, $t_{i,w}$ was set to a random time during the patient's ICU stay, and W was varied from 4 to 24 hours.

Features extracted from the window of fixed size W are detailed in Table 2. Features were extracted from a number of physiologic and laboratory measurements. Notably, no explicit data regarding treatment was extracted (e.g. use

of vasopressors, mechanical ventilation, dialysis, etc). Features extracted used consistent functional forms: the first, last, minimum, maximum, or sum (in the case of urine output) was extracted from all measurements of the variable made within the window W. As laboratory values are less frequently sampled, this window was extended 24 hours backward for these measurements. Table 2 lists the window sizes used for each of the features extracted. In addition to those listed in Table 2, a set of static features which do not change during an ICU stay were extracted. These included gender, age, ICU service, and whether the hospital admission was an emergency (binary covariate). Service was coded using one-hot encoding for the following services: coronary medicine, coronary surgery, surgery, neuro-surgery, other surgery, trauma, neuro-medicine, other medicine, orthopedic, gastro-utinery, gynecology, and ear/nose/throat. A total of 148 features were extracted.

Time window	Feature extracted	Variables
$[t_{i,w} - W, t_{i,w}]$	Minimum, Maximum, First, Last	Heart rate, Systolic/Diastolic/Mean blood pressure, Respiratory rate, Temperature, Oxygen Saturation, Glucose
$[t_{i,w} - W, t_{i,w}]$	Last	Glasgow coma scale, Glasgow coma scale components (motor, verbal, eyes), unable to collect verbal score
$[t_{i,w} - W, t_{i,w}]$	Minimum	Glasgow coma scale
$[t_{i,w} - W - 24, t_{i,w}]$	First, last	Base excess, Calcium, Carboxyhemoglobin, Methemoglobin, Partial pressure of oxygen, Partial pressure of carbon dioxide, pH, Ratio of partial pressure of oxygen to fraction of oxygen inspired, Total carbon dioxide concentration
$[t_{i,w} - W - 24, t_{i,w}]$	First, last	Anion gap, Albumin, Immature band forms, Bicarbonate, Bilirubin, Creatinine, Chloride, Hematocrit, Hemoglobin, Lactate, Platelet, Potassium, Partial thromboplastin time, International Normalized Ratio, Sodium, Blood urea nitrogen, White blood cell count
$[t_{i,w} - W, t_{i,w}]$	Sum	Urine output

Table 2: Features extracted during the window examined. Blood gases and laboratory measurements have the same feature extraction (first, last), but are separated for clarity. Note that some features have had their window extended backward an extra 24 hours.

Methods

We evaluated multiple machine learning models both to quantify the improvement possible from more flexible approaches and to ensure robustness of our findings across multiple settings. The models used were: logistic regression (LR), logistic regression with an L1 regularization penalty using the Least Absolute Shrinkage and Selection Operator (LASSO), logistic regression with an L2 regularization penalty (L2), and Gradient Boosting Decision Trees (GB).

In addition to the above models a set of five severity of illness scores were calculated on the data. Severity of illness scores are a form of mortality prediction model where the result is an integer score correlated with mortality. Severity of illness scores evaluated here are the APS III[4], SAPS[5], SAPS II[6], SOFA[7], LODS[8], and OASIS[9].

For all models, the outcome evaluated was in-hospital mortality. The area under the receiver operator characteristic curve (AUROC) was used to evaluate the models. The AUROC is the probability of ranking a patient who dies higher than a patient who lives: higher values of the AUROC indicate better model discrimination, and a value of 0.5 is equivalent to random chance. We also evaluated the models using the area under the precision recall curve (AUPRC), which provides a measure of performance which is agnostic to the number of true negatives and can be useful for problems with class imbalance. Hyperparameters were set using an internal three-fold cross-validation within the external five-fold cross-validation. Once hyperparameters were selected using the internal three-fold cross-validation, the model was retrained using the entire training set and subsequently evaluated on the external validation

fold. This process was repeated five times for each external validation fold. Hyperparameters searched were as follows: regularization penalty (LASSO/L2), the number of trees in the ensemble (GB), the learning rate (GB), and the maximum depth of each tree (GB). The other parameters were set to their default values. We used scikit learn v0.18.1[23] and XGBoost v0.6[24] with Python 2.7. Model performance is reported as the average across five held out folds in an outer cross-validation, with the minimum and maximum performance across all folds.

Experiments

Two experiments were conducted which essentially involve defining the window to be used for data extraction for each patient. In the first experiment, the window size was set to 24 hours ($W := 24$), and $t_{i,w}$ was fixed to 24 hours after ICU admission (i.e. $t_{i,w} := t_{i,adm} + W$). This emulates the most common framework for evaluating mortality prediction models used for risk adjustment. As these models aim to capture patient health on admission to the ICU (to reduce the impact of ICU care and acquire an estimate of how severely ill a patient was on admission), the data analyzed is from the first 24 hours (except labs, which are again allowed an additional 24 hours as in Table 2). We refer to this experiment as the "benchmarking experiment".

In the second experiment, $t_{i,w}$ is defined as a random time between ICU admission and ICU discharge or the first instance of a treatment limitation order, whichever is earlier, i.e. $t_{i,w} \sim \mathcal{U}[t_{i,adm}, min(t_{i,dis}, t_{i,dnr})]$. This experiment is designed to create a model which is applicable at all possible times during a patients ICU stay, rather than traditional models which are only applicable using data from the first 24 hours. Furthermore, in order to avoid the model learning from data corresponding to the withdrawal of treatment rather than severe illness, no data is extracted after a patient's code status is changed. For example, a do-not-resuscitate (DNR) order indicates that a patient does not desire to be resuscitated in the case of cardiac arrest. While some treatments are still given for patients who have a DNR, we took the conservative approach of excluding this data. Note that if a patient has a code status change, the data before this code status change would still be eligible for data extraction. Patients who had a code status other than full code on admission to the ICU were previously removed from this cohort as detailed earlier. We refer to experiment as the "real-time experiment". We utilized a window size $W := 4$ hours. We experimented with longer windows of 8 and 24 hours, though these did not impact model performance significantly (results not shown).

The models developed in the real-time experiment were also evaluated using windows ending at a fixed time to death. This experiment aims to evaluate the construct validity of the algorithms: as the window approaches patient mortality, we assume that the data will reflect worsening physiology. If the classifier has captured this information then the classifier's performance should increase. To this end, the models were trained as detailed above using a random time window during the patient's ICU stay. When evaluating the model, patients who died in the ICU had the end of the window set to 0, 4, 8, 16, and 24 hours before mortality. If the patient did not expire in the ICU, then a random time window during their ICU stay was used as was done during normal model development, and their outcome was set to zero (that is, a low prediction for a patient who dies more than 24 hours after ICU discharge is no longer penalized by the evaluation).

All code for this paper is open source and available online[25].

Results

Two exemplar patients are shown in Figure 1 with the two types of data extraction used for the benchmarking experiment and for the real-time experiment. The patient in the top plot survived to ICU discharge, while the patient in the bottom plot had a code status change and subsequently died in ICU. The gray background indicates the window used for data extraction, representing the benchmarking experiment in the top plot (first 24 hours of ICU admission) and the real-time experiment in the bottom plot (a random time during the ICU stay which cannot occur after a code status change).

Benchmarking experiment

The LR, LASSO, L2, and GB models were trained using features extracted as detailed for the benchmarking experiment. A total of 50,488 patients were included (the full cohort). Five fold cross-validation was used to obtain held

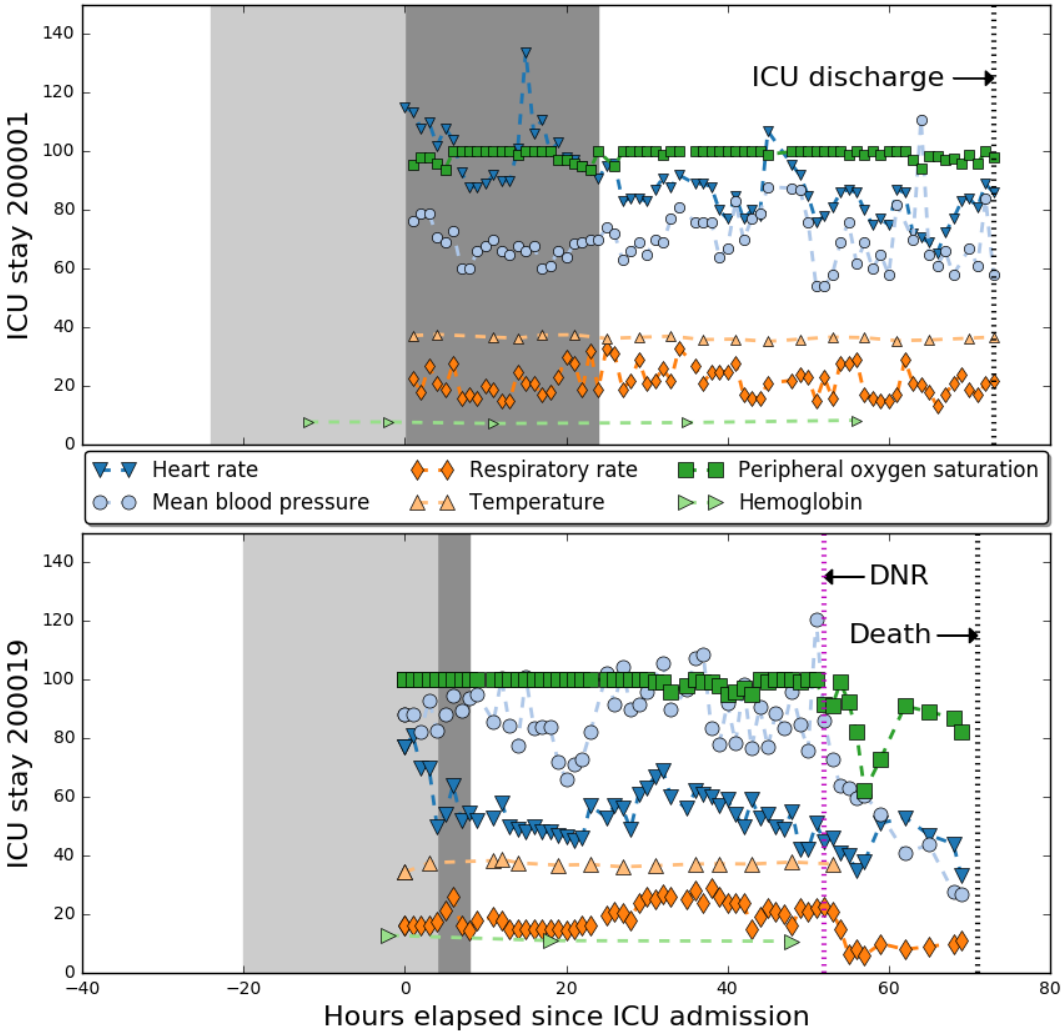

Figure 1: Two exemplar patients shown with vital sign data with measurement values along the y-axis. The shaded gray region corresponds to the window used for data extraction. In the top plot, the dark rectangular patch represents the window used in the benchmarking experiment (24 hours long), and the end of the window set to 24 hours after ICU admission. The lighter rectangular patch represents the additional window used for extracting laboratory measurements, which can be present before ICU admission. The patient in the top plot survived to ICU discharge. The bottom plot represents a distinct patient with the patches representing an example window (4 hours) used in the real-time experiment hours and the end of the window set to a random time during the patient's ICU stay before any code status changed. This patient was made do-not-resuscitate (DNR) and subsequently died in ICU.

out estimates of model performance. These models were compared to severity of illness scores: SAPS II, APS III, LODS, SOFA, and OASIS. AUROCs for severity of illness scores were calculated for each fold, and the minimum, maximum, and average AUROCs are reported, along with AUROCs for models trained in this work and models from the literature (Table 3).

	AUROC [minimum, maximum]	Cohort definition and size
SOFA	0.739 [0.735, 0.746]	*
LODS	0.755 [0.748, 0.760]	*
SAPS	0.758 [0.754, 0.765]	*
OASIS	0.774 [0.766, 0.780]	*
APS III	0.784 [0.774, 0.794]	*
SAPS II	0.809 [0.801, 0.822]	*
L2	0.897 [0.892, 0.899]	*
LASSO	0.892 [0.888, 0.897]	*
LR	0.896 [0.892, 0.899]	*
GB	**0.927 [0.925, 0.929]**	*
Lehman et al., 2012[18]	0.820	Exclude those missing SAPS-I, first ICU stay of at least 24 hours, N=14,739
Johnson et al., 2012[16]	0.860	First 48 hours in ICU, stayed at least 48 hours, random one third sample, N=4,000
Ghassemi et al., 2014[20]	0.840	First 12 hours in ICU, notes must contain 100 non-stop words, N=19,308
Caballero et al., 2016[21]	0.866	Random subsample, N=11,648
Luo et al., 2016[22]	0.848	First 12 hours, 30-day mortality

Table 3: Comparison of models trained here and in the literature with severity of illness scores. Models utilize data from the first 24 hours of a patient's ICU stay for predicting in-hospital mortality unless otherwise stated. For models trained in this work the average, minimum, and maximum performance across five folds of cross-validation is reported. *These models used all adult ICU admissions greater than 4 hours (N=50,488) split in 5-folds of cross-validation.

Real-time experiment

The real-time experiment assessed the performance of models both trained and evaluated using a random time point during the patient's ICU stay.

The performance of the LR, LASSO, RF, and GB models are shown in Table 4. These models were trained using data extracted from a window of length $W = 4$ hours located at a random time point during a patient's ICU stay, at least W hours before ICU discharge.

	AUROC [minimum, maximum]	AUPRC [minimum, maximum]
L2	0.892 [0.888, 0.896]	0.588 [0.568, 0.597]
LASSO	0.888 [0.882, 0.894]	0.579 [0.557, 0.593]
LR	0.892 [0.887, 0.896]	0.588 [0.568, 0.597]
GB	0.920 [0.918, 0.924]	0.665 [0.654, 0.669]

Table 4: AUROC and AUPRC of models trained and evaluated on distinct patients using a single window of data occurring at a random time during a patient's ICU stay (N=50,488). The average AUROC/AUPRC across 5-folds of cross-validation is reported, with the minimum and maximum value in brackets.

The models trained above were evaluated at fixed time points from patient mortality. Figure 2 shows the AUROC at different lead times from ICU discharge.

Figure 3 shows an example of applying the model to every hour of ICU stays for four patients; two of whom survived to hospital discharge and two of whom died in ICU.

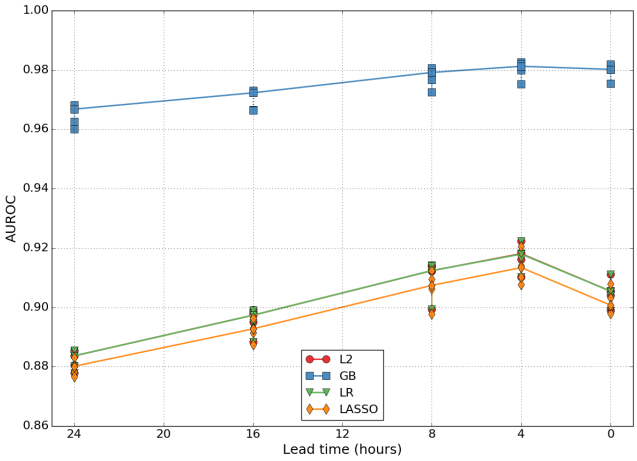

Figure 2: AUROC of models when $t_{i,w}$ is fixed at a certain distance from death for those patients who died in the ICU. If the patient survived past 24 hours from their ICU discharge, $t_{i,w}$ was set to random time during their ICU stay, and their outcome was set to 0 (i.e. they were treated as a survivor).

Discussion and Related Work

The benchmarking experiment demonstrates that the use of GB had a substantial improvement over previous methods. GB achieved an AUROC of 0.920, which is substantially higher than severity of illness scores evaluated on the same data (e.g. SAPS II with AUROC of 0.809). The improvement is likely attributable to a combination of more data and a more flexible modelling approach. The progression of the APACHE models for predicting mortality reflect this as well: later installments of the models had higher dimensionality and incorporated both cubic splines and interactions in order to better predict outcome[3]. The ability of GB to assign non-linear risk across values of a single feature could also contribute to its higher performance (e.g., low and high blood pressures could be assigned distinct risk estimates).

Interestingly, the LR model was extremely competitive, achieving an AUROC of 0.892. The LASSO model was slightly worse than LR (AUROC=0.888). In medicine, a key component of any decision support tool is interpretability, and many practitioners will not trust a tool if they do not understand how it produces predictions. While higher performance is certainly a desirable aspect of the model, the use of a simpler regression model which can be interpreted still remains a promising avenue of future research.

Comparing this performance to the literature, we see that the GB had a slightly higher AUROC than published for the APACHE IV model (AUROC of 0.89)[3]. It is worth noting that the models here used purely physiology, laboratory measurements, and minimal demographic variables available on admission, whereas the APACHE system utilized over 100 diagnostic categories, comorbid burden, and treatment. These additional features can often require manual collection which can complicate automatic application of the model. In a comparison of APACHE IV versus SAPS II, data abstraction per patient took much longer for APACHE IV (37.3 minutes) versus SAPS II (19.6 minutes)[26], and the difficulty in acquiring features was certainly a large factor there. The use of automatically or routinely collected features has many advantages in the potential practical application of models such as the one presented here. Further, the capability of the models to predict well given only physiology may indicate that with enough measurements regarding the patient manually laborious and potentially ambiguous tasks such as classifying patients into a single diagnosis could be circumvented.

The models developed here resulted in higher AUROCs than those reported in the literature. While it could be claimed that the results here represent the "state of the art" in mortality prediction for ICU patients, there are too many caveats to make this statement. First, we cannot claim to have exhaustively located all mortality prediction models built using the MIMIC database or otherwise. However, more importantly, we cannot even claim that the GB model is the best

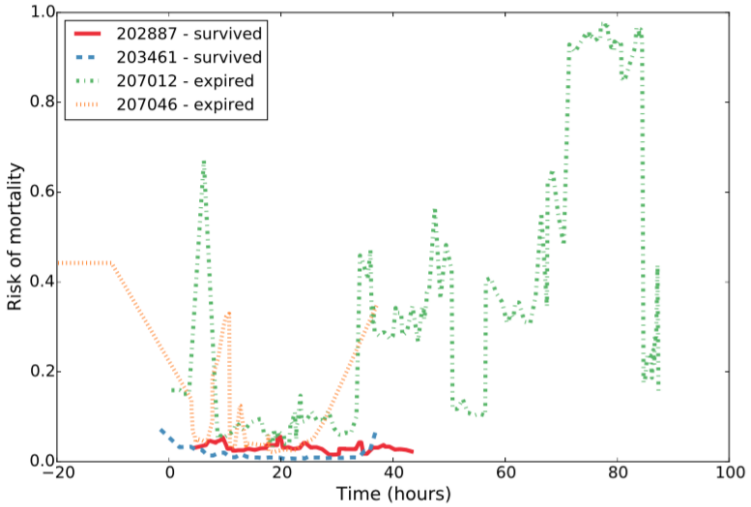

Figure 3: Example trends for four patients in the ICU. Two of the patients survived to ICU discharge, while two of them died in the ICU.

of those reported in Table 3, even though all utilized some version of the MIMIC database. This is due to a number of reasons. First, there is a lack of consistency in the dataset definition and extraction, and in particular in the timing of data extraction. While most studies utilize the first 24 hours, some required the patients to stay at least 24 hours, while others utilized 48 hours and required patients to stay 48 hours. These decisions substantially impact the patient case-mix and consequently the difficulty of the problem. This is even highlighted by Luo et al.[22], who utilize the first 12 hours of mortality as it is a more challenging prediction task. In addition to timing, studies frequently disregard samples due to missing data[18,19], inappropriate patients (this study removed organ donor accounts), and in order to reduce heterogeneity in the cohort (Hug et al.[14,22] removed neuro-surgical and trauma patients).

As MIMIC is a completely open database, and as tools to facilitate sharing of data analysis have matured, there is a unique opportunity for research in this area to be fully reproducible. We hope that future work will aim to alleviate this issue by releasing source code associated with their analysis. In this spirit, we have made all code for this study public and hope that this facilitates future research to improve upon the benchmark we have established[25].

The real-time experiment was consistent with the benchmarking experiment in that the best performing model was GB (AUROC = 0.920), with the LR model having competitive performance (AUROC = 0.892). The AUROC of the LR model was slightly lower than the model using data from the first 24 hours (0.896), though not greatly. The L2 model had approximately equivalent performance to the LR model in all experiments.

Figure 2 demonstrates that as the models approach the time of mortality their ability to discriminate mortality increases. This is intuitively sensible and acts as a sanity check that the models are capturing abnormal physiology which relates to mortality, though it is not clear why the GB model had a much lower improvement as opposed to other models.

In the medical domain it is important to consider the interpretability of the final model developed. In Figure 3 we see an example of how the predictions of the model would result in a trend, as if used at the bedside continuously, and there appear a few spikes in the predictions. The use of a LR model would allow easy explanation of why the risk profile changed: the following covariates changed which consequently increased or decreased the final prediction. However, a similar interpretation cannot easily be produced using GB, as multiple features will have changed and translating the impact of these changes within the model is non-trivial. The interpretation of tree based models continues to be an active area of research. The trends shown in Figure 3 also show that there is a great deal of information in the trend which could be used to further improve the model performance.

There are limitations to our study. First, as all data in MIMIC is acquired from a single hospital, these models may not be applicable for outside datasets. However, the applied methodology is relatively simple, and we plan to evaluate a

similar framework in a large multi-center database. Second, the exact mechanism for utilizing such a model is unclear. While synthesizing patient health as presented could have applications for resource utilization and decision support, integrating such a model into clinical practice is a difficult process which must be made with careful consideration. Finally, the data used was a small subset of what is available in MIMIC. Specifically, we did not utilize any data regarding patient treatment, high resolution waveforms, or notes written during routine patient care. The incorporation of information from these sources may improve model performance.

Conclusion

We have shown that classic mortality prediction models can be improved by the use of flexible machine learning approaches combined with more granular data. A GB model had the highest performance here and than models reported in the literature, but we argue that the field must adopt benchmark datasets and a dedication to making code openly available code in order to both validate this claim and to allow research to progress in this field. Finally, we demonstrated a feasible architecture for developing a real-time mortality prediction system for evaluating patient health. The predictions made may provide clinicians with an accurate and rapid summary of patient health, though further research is necessary in order to ensure models are both interpretable and usable at the bedside.

References

1. R L Kane, T A Shamliyan, C Mueller, S Duval, and T J Wilt. The association of registered nurse staffing levels and patient outcomes: Systematic review and meta-analysis. *Medical Care*, 45(12):1195–1204, December 2007.

2. W A Knaus, J E Zimmerman, D P Wagner, E A Draper, and D E Lawrence. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Critical Care Medicine*, 9:591–597, 1981.

3. J E Zimmerman and A A Kramer. Outcome prediction in critical care: the Acute Physiology and Chronic Health Evaluation models. *Current Opinion in Critical Care*, 14:491–497, 2008.

4. W A Knaus, D P Wagner, E A Draper, J E Zimmerman, M Bergner, C A Bastos, P G snd Sirio, D J Murphy, T Lotring, A Damiano, and F E Harrell Jr. The APACHE III prognostic system: Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100(6):1619–1636, 1991.

5. J R Le Gall, P Loirat, F Nicolas, C Granthil, F Wattel, R Thomas, P Glaser, P Mercier, J Latournerie, P Candau, et al. Use of a severity index in 8 multidisciplinary resuscitation centers. *Presse medicale (Paris, France: 1983)*, 12(28):1757–1761, 1983.

6. J R Le Gall, S Lemeshow, and F Saulnier. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA*, 270(24):2957–63, 1993.

7. J-L Vincent, R Moreno, J Takala, S Willats, A De Mendoca, H Bruining, C K Reinhart, P M Suter, and L G Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22:707–710, 1996.

8. Jean Roger Le Gall, Janelle Klar, Stanley Lemeshow, Fabienne Saulnier, Corinne Alberti, Antonio Artigas, and Daniel Teres. The logistic organ dysfunction system: a new way to assess organ dysfunction in the intensive care unit. *Jama*, 276(10):802–810, 1996.

9. A. E. W. Johnson, Andrew A Kramer, and Gari D Clifford. A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. *Critical Care Medicine*, 41(7):1711–1718, 2013.

10. K Strand and H Flaatten. Severity scoring in the icu: a review. *Acta Anaesthesiologica Scandinavica*, 52(4):467–478, 2008.

11. Mark T Keegan, Ognjen Gajic, and Bekele Afessa. Severity of illness scoring systems in the intensive care unit. *Critical care medicine*, 39(1):163–169, 2011.

12. William A Knaus. Apache 1978-2001: the development of a quality assurance system based on prognosis: milestones and personal reflections. *Archives of Surgery*, 137(1):37–41, 2002.

13. Alistair E W Johnson, Tom J Pollard, Li-wei Lehman Lu Shen, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 35, 2016.

14. Caleb W Hug and Peter Szolovits. Icu acuity: real-time models versus daily models. In *AMIA*, 2009.

15. Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *Computing in Cardiology (CinC), 2012*, pages 245–248. IEEE, 2012.

16. Alistair E W Johnson, Nic Dunkley, Louis Mayaud, Athanasios Tsanas, Andrew A Kramer, and Gari D Clifford. Patient specific predictions in the intensive care unit using a bayesian ensemble. In *Computing in Cardiology (CinC), 2012*, pages 249–252. IEEE, 2012.

17. Luca Citi and Riccardo Barbieri. Physionet 2012 challenge: Predicting mortality of icu patients using a cascaded svm-glm paradigm. In *Computing in Cardiology (CinC), 2012*, pages 257–260. IEEE, 2012.

18. Li-Wei H Lehman, Mohammed Saeed, William J Long, Joon Lee, and Roger G Mark. Risk stratification of icu patients using topic models inferred from unstructured progress notes. In *AMIA*. Citeseer, 2012.

19. Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84. ACM, 2014.

20. Marzyeh Ghassemi, Marco A F Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2015, page 446. NIH Public Access, 2015.

21. Karla L Caballero Barajas and Ram Akella. Dynamically modeling patient's health state from electronic medical records: A time series approach. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 69–78. ACM, 2015.

22. Yuan Luo, Yu Xin, Rohit Joshi, Leo Celi, and Peter Szolovits. Predicting icu mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements. In *AAAI*, pages 42–50, 2016.

23. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

24. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*, 2016.

25. Alistair E. W. Johnson. alistairewj/mortality-prediction. doi:10.5281/zenodo.823359. Jul 2017.

26. Michael W Kuzniewicz, Eduard E Vasilevskis, Rondall Lane, Mitzi L Dean, Nisha G Trivedi, Deborah J Rennie, Ted Clay, Pamela L Kotler, and R Adams Dudley. Variation in icu risk-adjusted mortality: impact of methods of assessment and potential confounders. *CHEST Journal*, 133(6):1319–1327, 2008.