Feature Representation for Predicting ICU Mortality

Harini Suresh, Yun Liu, Collin Stultz

***Abstract*—Effective predictors of Intensive Care Unit (ICU) Mortality have the potential to identify high-risk patients earlier, improve ICU resource allocation, and create more accurate population-level risk models. Machine learning practitioners typically make choices about how to represent features in a particular model, but these choices are seldom evaluated quantitatively. This study compares the performance of different representations of clinical event counts from the MIMIC III database in a logistic regression model to predict post-36-hour ICU mortality. The most common representations are linear (normalized counts) and binary (yes/no). These, along with a logarithmic representation and a new representation termed Hill, are compared using L2 regularization. Results indicate that the introduced Hill representation gives a higher area under the reciever operating curve (AUC) for the mortality prediction task than the log, binary and linear representations. The Hill representation thus has the potential to improve existing models of ICU mortality.**

***Index Terms*—ICU, Mortality, Prediction, Data, Logistic Regression, Prognosis**

Introduction and Related Work

PROGNOSTIC models to predict the outcome of patients in Intensive Care Units (ICU) are valuable for many reasons, among them:

1) Risk stratification: stratifying patients by their risk to various adverse events provides a way to evaluate and compare ICUs and new therapies. For example, if one hospital has a higher mortality rate than another, it does not necessarily mean that the hospital is performing more poorly. It may just be a reflection of a difference in the average health of the two different patient populations. The ability to empirically risk stratify patients essentially allows these evaluations calibrate themselves to the unique state of patients in the hospital for more accurate comparisons [25].
2) Optimizing resource utilization: the ICU is a high-cost, emotional, and resource-constrained environment. ICUs are already over-crowded, and many patients are not able to receive critical care that would be beneficial to them [9]. In this environment, utilization strategies are clearly essential. Detsky et. al. showed that both total expenditure and expenditure per day in the ICU were highest for patients whose outcomes were the most unexpected when compared to a physicians predicted prognosis [4]. Being able to predict how at-risk various patients are throughout their stay provides an empirical basis for scheduling and resource allocation. It also provide estimates for how long a patient should continue a therapy or what the optimal time for discharging a patient is [11].
3) Clinical decision-making: predictive models can provide a reliable and unbiased way to use past experiences to guide future ones. Besides reducing expenditures [4], this guidance would lead to more efficient and helpful care for patients [6]. Physicians perform clinical decision-making everyday. However, a data-derived prognostic model provides the advantage of being supported by more data than any one physician's experiences as well as being certainly unbiased. Studies have shown that when implemented effectively, prognostic models have improved patient care. For example, the Thrombolytic Predictive Instrument (TPI) estimates the risk of key outcomes of thrombolytic therapy. Sekler et. al. performed a randomized controlled clinical effectiveness trial and showed that printing the TPI on electrocardiogram headers improved and expedited the appropriate use of therapies for patients [24].

Currently, the most widely used methods of mortality prediction in clinical practice are the most recent versions of the Simplified Acute Physiology Score (SAPS II) and Acute Physiology and Chronic Health Evaluation (APACHE II). The SAPS II score is calculated from 12 physiological variables (such as age, heart rate, and blood pressure) and 3 disease-related variables (such as chronic diseases present and type of admission) measured during the first 24-hours of admission [14], similar to the APACHE II score [10]. Both scores use logistic regression models to predict in-hospital mortality. However, many evaluations of these scores have shown that are not able to achieve adequate calibration (that is, the ability to provide a risk estimate corresponding to observed patient mortality) [1], [17], [21]. Studies have shown that their performance in mortality prediction tasks is not accurate enough to provide a significant advantage over the physician's own prognosis [28]. However, improvements in performance have been made by augmenting these models with additional variables [1], [14], [15], [16].

Since 2001, the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) database has been built up and maintained by the Laboratory of Computational Physiology at the Massachusetts Institute of Technology, Beth Israel Deaconess Medical Center, and Philips Healthcare, with support from the National Institute of Biomedical Imaging and Bioinformatics [23]. The most recent version of this database, MIMIC III, contains data from around 38,600 adults, comprising over 58,000 hospital admissions, from 2001-2012. The data includes features such as demographics, bedside vital sign measurements, laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of hospital). MIMIC is unique in its scale, as well as the robustness of the included variables and inclusion of highly granular data.

Some studies have attempted to augment basic logistic regression mortality predictors like SAPS and APACHE with the great deal of additional data provided in MIMIC [3].

Incorporating this additional information can aid in creating more patient-specific models of mortality [26]. Studies have found that these more robust models perform significantly better at mortality prediction tasks for diverse subsets of patients than the existing scoring systems [2].

There are several different ways to interpret the features themselves. In the MIMIC database, many features are numerical values (i.e. white blood cell count). When using these in a model, there is a choice of whether to count simply the number of times the clinical event occurs (i.e. the number of times a patient's white blood cell count was measured) versus the actual values of the feature. Using the values is considered very patient-centric, since it revolves around exactly what is happening to the patient. Using counts, on the other hand, is more physician centric, since it reflects the physician's actions and what he or she actually decided to do for the patient. Models using either interpretation of the data capture different aspects of the patient's condition, and therefore can both be valuable, since the predicted prognosis is one aspect incorporated into a larger holistic understanding of patient's status. In this study, we use counts of clinical events for a more physician-centric model, similar to some previously effective studies [18], [29] .

When building these models, machine learning practitioners typically choose numerical feature representations based on the type and scale of the features in the data. Speaking generally, in machine learning, giving attention to feature representation provides many advantages, such as improving the performance of predictions and providing a better understanding of processes underlying the data [8]. Functions of features that are well-conceived can take into account the inherent properties and distributions of the different features and capture important information in data better than the raw values of the variables would.

In the mortality prediction task, the most prevalent representations are binary (yes/no) [12] or linear (normalization based on min/max values) [18]. However, these representations are not usually adequate. A binary representation may discard information about the severity of a patient's condition since it treats all positive values as the same (i.e. a patient receiving multiple doses of drugs promoting urination may be having more trouble with water retention, but this would not be reflected in a binary representation). On the other hand, a linear representation may grant this information too much importance, since it assumes that severity increases linearly with the feature value (i.e. a patient receiving three types of blood pressure lowering medications may not be at 3 times the risk of cardiovascular events relative to a patient receiving one). Additionally, linear representations assume proportional increases in severity no matter where a value lies, giving disproportionate importance to outliers.

Despite the weakness in these representations, these choices are rarely evaluated quantitatively or compared with other possibilities. Other possibilities include a logarithmic representation and a new approach that will be introduced as the Hill representation, which is centered around a feature's median value. Different representations are also likely better suited to different features, due to the different inherent distributions and properties of the features. For example, for one particular feature, simply its presence may convey all the information that is needed, implying that a binary representation would be best. For another, there might be various amount of information conveyed in increasing values, and this information may be different depending on how close the values are to a median or average. A more complex representation would be better suited for this feature.

Introducing new feature transformations also allows for some nonlinearity in models such as logistic regression which are otherwise constraised by linear and additive relationships between features and the outcome. While logistic regression models are the most direct extension of existing predictive models (i.e. SAPS and APACHE), stringent linearity may not be a realistic assumption given the complex processes underlying ICU mortality [20].

This study aims to quantitatively evaluate the strengths and weaknesses of many different feature representations in the mortality prediction task, with the long-term goal of introducing a method to choose optimal representations for each feature, from among a wide library of representations. If, from a given set of selected features, the best representations could be successfully determined, over/under-representation problems could be minimized to give the most accurate model of ICU mortality. This in turn has the potential to better treatment of ICU patients as well improve accuracy of population-level risk models, among many other beneficial implications.

METHODS

Features

Features were extracted from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC III) Database [23]. MIMIC III contains data collected between 2001 and 2012 from a variety of ICUs in a single tertiary teaching hospital. The database contains general patient information (ICD-9 codes, demographics, room tracking), physiological signals (vital metrics, SAPS), medications (IV meds, provider order entry data), lab tests (chemistry, imaging), fluid balance (intake, output), and notes (discharge summary, nursing progress reports). The MIMIC dataset is notable because it is publicly available for free use, encompasses a large and diverse set of patients, and contains numerous high resolution features for each patient.

The study used data from patients in the Medical Care Unit (MICU), Cardiac Care Unit (CCU), Cardiovascular Intensive Care Unit (CVICU), Medical/Surgical Intensive Care Unit (MSICU), Surgical Intensive Care Unit (SICU), and Trauma Surgical Intensive Care Unit (TSICU). The features extracted were from the following MIMIC tables: patients, chartevents, inputevents, outputevents, microbiologyevents, procedureevents, and prescriptions (see table I).

All clinical event data was represented as a count (for example, in the case of a procedure, the feature value is how many times that procedure was administered, rather than any values associated with it). As described in the previous section, using counts leads to a more physician-centric model.

MIMIC Table	*Type of Data*	*Example*
patients	demographics	gender, age
chartevents	vital signs, patient status	sitter at bedside, hemoglobin
inputevents	fluids or medicines given to patient	cyclosporin
outputevents	fluids leaving the patient	vomit
microbiologyevents	lab tests, sensitivies	streptococcus species
procedureevents	procedures performed	bile duct repair
prescriptions	prescribed drugs	heparin

TABLE I: Examples of the information contained in the MIMIC III tables that were used to build the model.

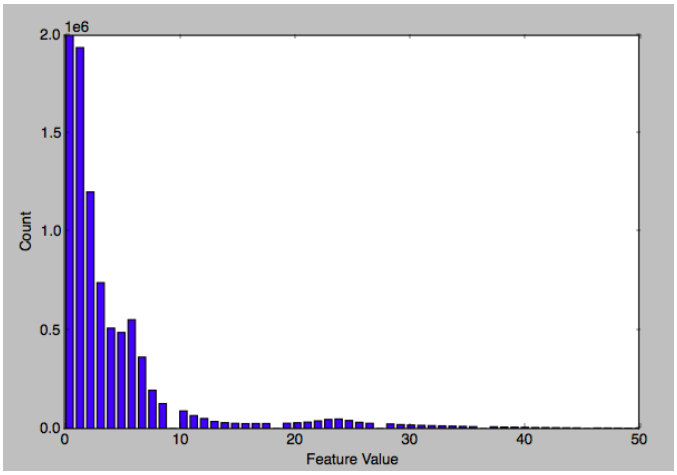

Fig. 1: The distribution of extracted feature values. Most values are 0 or 1, but there is a tail of larger values. Since the data contains values on such different scales, it is necessary to transform them somehow so one feature is not operating on a vastly different range than another. The question we ask is how to transform these larger values before using them to build a model.

This data comprises approximately 32,000 patients and 8,000 features. Figure 1 contains the distribution of feature values of the extracted features. Most of the features are zero or one, but there is a signficant tail of larger values. The existence of this tail indicates that features are distributed on different scales, and some kind of transformation is needed. The representation question asks how to transform these other values before feeding them to the model.

Prediction Task

Like the majority of existing mortality prediction models, we use features extracted from the first 24 hours of a patient's stay. This makes our model easy to compare to others like SAPS and APACHE, and also ensures that it is realistically usable. In a clinical use case, physicians would likely want a prognosis prediction early in the patient's stay (i.e. 24 hours in). In that situtation, we would also not have any future information past the 24-hour mark, so we take care not to use future information which would not be available in a real-time siutation. This data is used to predict mortality at any time from 36 hours after admittance to the end of the patient's stay (figure 2). We choose to begin the prediction task after

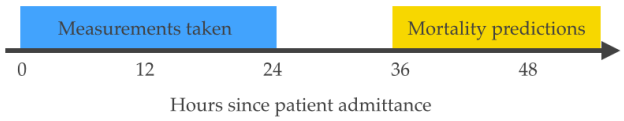

Fig. 2: Features were extracted from the first 24 hours of a patient's stay, and the prediction task begins 36 hours after admittance. This replicates a real-time situation where we would need to make predictions based only off of past data.

36 hours since a mortality between 24 and 36 hours is likely less of an indicator of the strength of the model and more an indicator of an inevitable outcome of the patient. In other words, if a patient will die in the 12 hours, it is highly likely that the physicians are already aware that the patient is at high risk.

Representations

We can describe the various representations as mathematical functions of the raw feature value. In the following equations, the raw feature value for a patient i and feature j is indcated with z_{ij}, and its transformed representation in the feature matrix is x_{ij} (see figure 3 for graphical versions of each representation).

The first and simplest representation is the Binary representation, or the indicator function for a non-zero value of z_{ij}:

$$binary\{x_{ij}\} = I(z_{ij} > 0) \tag{1}$$

An alternative representation that preserves the severity information contained in different non-zero raw feature values is a linear map of z_{ij} to the range [0,1]:

$$linear\{x_{ij}\} = \frac{z_{ij} - min_j}{max_j - min_j} \tag{2}$$

The linear representation can also be truncated or untruncated, a distinction that comes about if the testing set has a different range than the training set. If the testing set has a value that is past the maximum of the training set, the testing set representation can either be cut off at 1, or scaled up proportionately (see figure 4).

A logarithmic representation was also tested, where the values were normalized in the range [0,1] after the log-transform:

$$log\{x_{ij}\} = \frac{log(z_{ij}) - min_j}{max_j - min_j} \tag{3}$$

Care Unit	*Number of Patients*	*Number of Patients with Outcome of Mortality*	*% of Patients with Outcome of Mortality*
Cardiac Care Unit (CCU)	4689	476	10.1
Cardiovascular Intensive Care Unit (CVICU)	6865	203	3.0
Medical Care Unit (MICU)	6473	932	14.4
Medical/Surgical Intensive Care Unit (MSICU)	4224	571	13.5
Surgical Intensive Care Unit (SICU)	5080	586	11.5
Trauma Surgical Intensive Care Unit (TSICU)	4217	385	9.1

TABLE II: Patient outcome proportions in each of the care units used in the study.

Finally, a novel representation similar to the Hill equation in biochemistry is as follows:

$$hill\{x_{ij}\} = \frac{z_{ij}}{z_{ij} + m_j} \tag{4}$$

where m_j is the median value of non-zero z_{ij} across all patients feature j. The Hill representation is drawn from the Hill equation in biochemistry, which quantifies the binding of a ligand to a macromolecule given the existing ligand concentration. The idea that is captured in this equation, and carried over to its use as a feature representation, is that incremental increases of identical size carry different amounts of importance based on where in the range of values they fall. In biochemistry, it often takes much less energy for a ligand to bind to a macromolecule if there are already other ligands bound to it (i.e. cooperativity), and more energy if the ligand is the first to bind to the macromolecule. In a parallel sense, an increase from 0 to 1 prescriptions of a certain medicine may indicate more of an increase in risk than an increase from 18 to 19 (at that point, we may already know that the patients are at high risk).

In this representation, values equal to the median are mapped to 0.5. Using the median allows for less sensitivity to outliers, and allows the representation to be independent of the feature's minimum or maximum. It also provides greater resolution around the range of most common values.

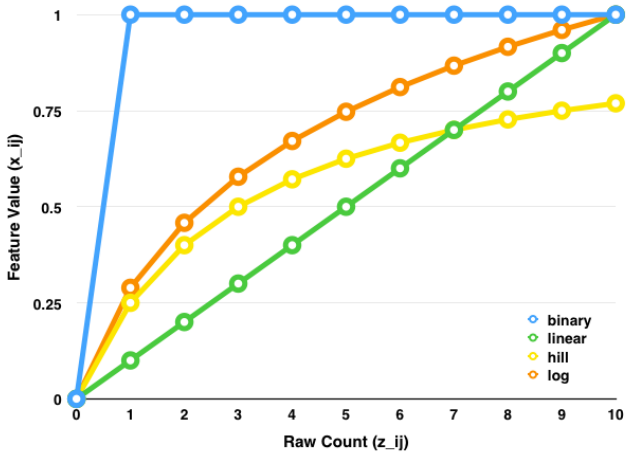

Fig. 3: A schematic of the different feature representations for the values 1 through 10 for a feature with a minimum of 0, maximum of 10, and median of 3. Note that the Hill representation has a transformed value equal to 0.5 at the median.

Model

A logistic regression model with L2 regularization was used for prediction. Logistic regression attempts to find $\hat{\boldsymbol{\theta}}$ (which corresponds to the coefficients of feature values in a logistic function) that maximizes the following likelihood function:

$$\hat{\boldsymbol{\theta}} = argmax_\theta \sum_{i=1}^{m} \log p(y_i|x_i, \theta) - \lambda R(\theta) \tag{5}$$

The $R(\theta)$ term is the regularization term, which attempts to combat overfitting by driving the θ values closer to zero. The constant λ is a cost parameter that scales the regularization term. In L1 regularization, $R(\theta)$ is the sum of the absolute values of each θ_i in θ. In L2 regularization, $R(\theta)$ is the sum of the squared values. Since the term is negated in the maximization, we are trying to minimize $R(\theta)$. Minimizing the absolute, rather than the squared, values of the parameters

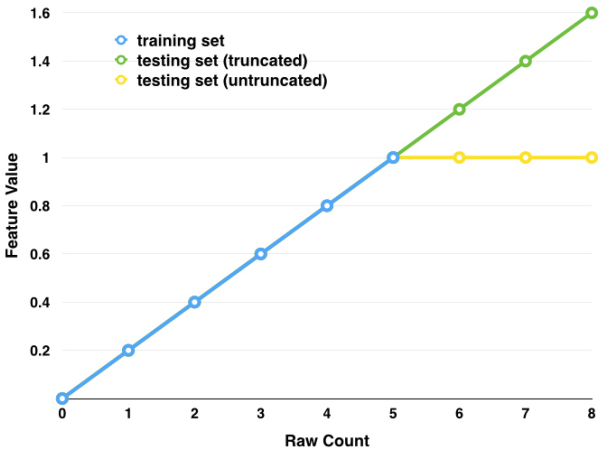

Fig. 4: A schematic of a truncated versus non truncated linear representation, using a dataset where the training maximum is 5, but values larger than 5 appear in the testing set.

drives the resulting coefficients (contained in $\hat{\theta}$) down to zero more drastically. Both L1 and L2 regularization were tested in preliminary experiments, but L2 consistently performed better. This is likely due to the fact that L1 reduces dimensionality by driving more coefficients to zero, however, in these experiments the marjority of features that are used are relevant, and benefit from having nonzero coefficients. Therefore, the models that were compared use L2 regression.

The logistic regression implementation is from sklearn's linear_model.LogisticRegression module. The training and testing sets were split using a stratified 80/20 ratio, and the train/test split was repeated 10 times for each transformation with different random seeds. 3-fold stratified cross-validation was used on each training set to determine the optimum cost parameter for each model.

Stratified sampling was used to split the training and testing sets and to do cross-validation since the dataset exhibits class imbalance. In other words, the relative proportion of patients who died is low, so it is possible that with random sampling the training or testing set could end up with a disproportionately low percentage or no patients with this outcome. See table II for more detailed patient proportions in the various care units.

RESULTS

After training each model on the training set, the models predicted mortalities for the testing set.

Area Under the Curve

The predictions were evaluted using the area under the receiver operating characteristic (ROC) curve, or area under the curve (AUC). An ROC curve is a common way to visualize the performance of a binary classifier. The y-axis of an ROC curve is the true positive rate (TPR) of predictions, and the x-axis is the false positive rate (FPR).

A logistic regression model outputs probabilities for a given outcome (i.e. mortality). In order to use these probabilities to perform classification (i.e. predict whether or not a patient will die, rather than the probability that they will), we set a probability threshold, and say that if a patient's probability of mortality is above the threshold, we predict mortality. The ROC curve plots the TPR versus FPR for all possible thresholds we could choose. This leads to much more robust evaluation metric than simply using the error rate for a single threshold.

The best point on the plot would be in the upper left corner, corresponding to a TPR of one and FPR of zero. The worst points are those where the FPR is equal to or greater than the TPR, since this is equal to or worse than random guessing. The closer a point is to the top left corner of the ROC curve (and consequently, the greater the area under the curve), the better the predictor is at discriminating between the two classes. By maximizing the AUC, we also take into account the model's ability to discriminate, rather than just its accuracy alone (for example, a model that predicted all patients would live would have relatively good accuracy, since a small percentage of patients acutally die, but this model would not be good at separating classes).

Model Comparison

AUCs were measured and averaged across all 10 train/test splits for a particular transformation (see fig 5). The Hill representation gave the highest average AUC, followed by binary and log (in some care units, binary performed better than log, and in others vice versa), truncated linear, raw values, and untruncated linear. To assess significance, the Hill representation AUC was compared to the other AUCs for each train/test split. In all care units except for CVICU, the Hill representation model resulted in a higher AUC for at least nine out of ten train/test splits.

DISCUSSION

Model Performance

The Hill representation led to a better average AUC than linear, binary, and log values for predictions all care units except CVICU. It's better performance may be because the linear representation uses the maximum function, which is sensitive to outliers. Specifically, when some patients in the population have a high count of a feature, the bulk of the population feature values get compressed by the comparatively large value in the denominator of Equation 2. By contrast, the Hill representation spreads the density more evenly. On the other hand, the binary representation loses all granularity within positive values of a particular count. The Hill representation usually performs better than the logarithmic representation as well, indicating that there is importance in 1) having high resoulution specifically around the median, and 2) not tapering off increasing values to the extent that the log transform does.

The models run on the CVICU care unit exhibit a slightly different behavior in the performance of the different representations (it does not take on the order described above and seen in the other care units, with Hill giving the highest AUC followed by log/binary and then linear). This may be because the percentage of patients with an outcome of mortality in this dataset is much lower than in the data for the other care units (see table II). Given this much smaller proportion of one class, the models for this care unit may not have been trained to the same extent as the others, and therefore behave slightly differently.

In general, the AUCs of the models in this study were slightly lower than the state of the art models [20] for morality prediction using MIMIC, which are able to achieve AUCs of around 0.88 across the board. This may be largely due to the fact that this study uses counts of clinical events rather than their actual numerical values. While using counts captures a unique physician-centric aspect of the patient's condition, we lose some of the information contained in the numerical values of these features that may lead to more accurate mortality predictions.

Top Features

An example of the top features for the model trained using all care units and the Hill feature transformation is in table III. The feature with the highest weight is the SAPS score, which is likely because it is an already formulated severity score

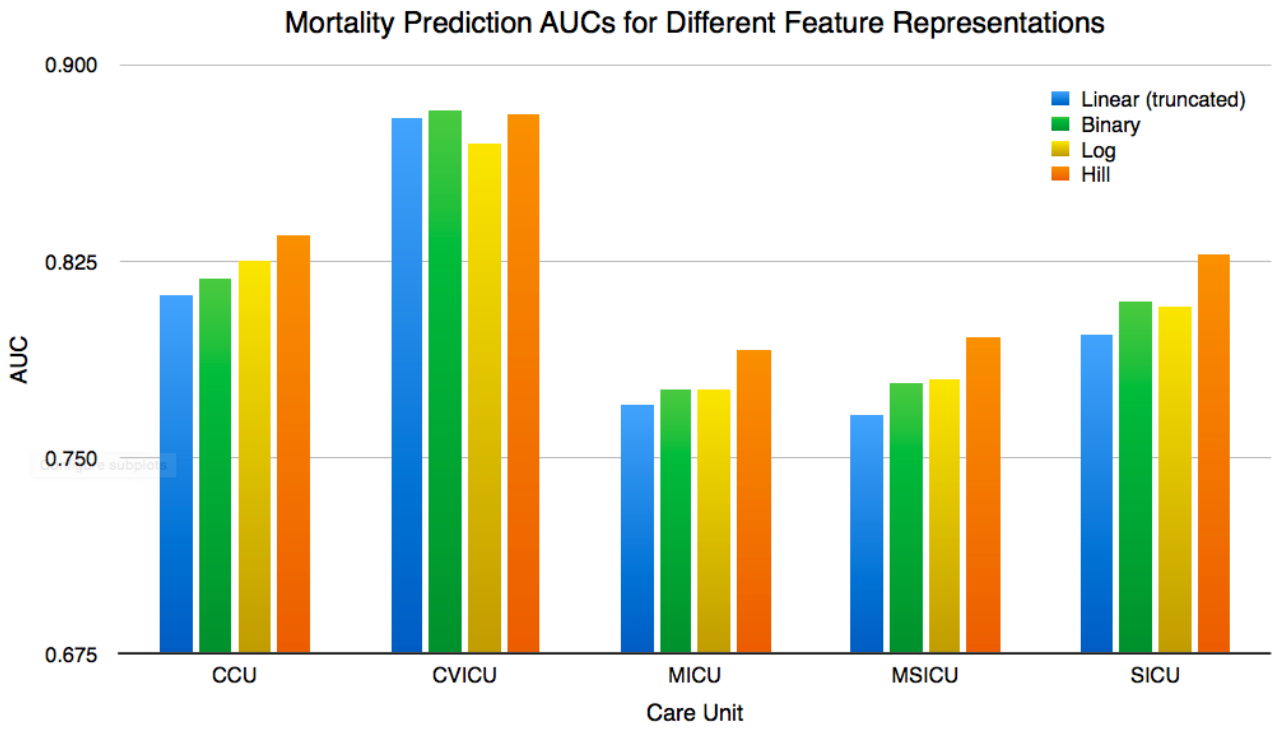

Fig. 5: The AUCs of each model on the testing set, for each care unit on its own (the model was trained separately for each care unit, using training data from just that care unit). Raw values and untruncated linear representations performed significantly worse than the others and are not included in the figure. The average AUC of the Hill representation model is higher than the other representation models in all care units except for CVICU.

Feature	*Weight*
SAPS-I Score	2.01933
Endoscopic retrograde cholangiopancreatography (procedure)	0.713268
Aprotinine	-0.707276
Incision of perirenal or periureteral tissue (procedure)	0.620437
Operation on thorax (procedure)	-0.604544
Partial hepatectomy (procedure)	0.484179
Caval pulmonary artery anastomosis (procedure)	0.475672
Bile duct repair (procedure)	0.466424
Percutaneous biliary drainage (procedure)	0.464564
Disconnect alarm	-0.450901

TABLE III: The features with the highest absolute weights in the L2 Logistic Regression Model with features transformed with the Hill transformation and trained on all care units. Positive weights indicate factors that push the predicted mortality probability closer to one (mortality) and negative weights indicate factors that push the prediction closer to zero (survival).

that incorporates and scales various important indicators of risk. Because it is a severity score, a higher score corresponds to more risk, leading to the positive coefficient. The next feature, endoscopic retrograde cholangiopancreatography (a procedure) also has a positive coefficient, indicating that its presence increases risk. This is a procedure which treats problems of the bile and pancreatic ducts, and is most commonly performed on patients who have cancer that has advanced into a tumor that narrows or blocks bile or pancreatic ducts. The fact that patients who receive this treatment are of a more at-risk population may lead to its high coefficient.

The majority of other important features are also procedures or procedure-related. Some of these also have positive weights, which could be attributed to a reason similar to that described previously, related to the other likely complications of patients who must undergo that procedure. For example, a partial hepatectomy, the eighth feature on the list, is a procedure that involves removing part of the liver due to a metastasized tumor. These positive weights can also be due to the riskiness of the procedures. For example, bile duct repair, the tenth feature on the list, is considered a technically challenging procedure, and injuries to the bile duct can triple morbidity [7].

A few of the other top features have negative weights, indicating that their presence leads to a lower risk prediction.

Aprotinine, for example, is a widely-used drug that reduces bleeding and the need for blood transfusions. As a result, it is likely used during many procedures that are routinely performed, not very risky, or not associated with especially severe patient conditions.

Future Work

In the future it may be interesting to include new types of representations, including a "discrete" representation where features are assigned a value based on certain quantiles of the feature distribution rather than the actual minimum and maximum. While L2 regularization was used in the results presented in this study, it may also be beneficial to try different regularization methods such as elastic net, a balance between L1 and L2 regularization. In addition, switching to using numerical values of features rather than counts would likely lead to better performances by the models.

Finally, a long term goal of studying feature representations is to develop unique representations for each individual feature that reflect the specific properties and distribution of that feature.

REFERENCES

[1] Arabi, Yaseen, et al. "Assessment of six mortality prediction models in patients admitted with severe sepsis and septic shock to the intensive care unit: a prospective cohort study." *Critical care* 7.5 (2003): R116.

[2] Celi, Leo Anthony, et al. "A database-driven decision support system: customized mortality prediction." *Journal of personalized medicine* 2.4 (2012): 138-148.

[3] Anthony Celi, Leo, et al. "Big data in the intensive care unit. Closing the data loop." *American journal of respiratory and critical care medicine* 187.11 (2013): 1157-1160.

[4] Detsky, Allan S., et al. "Prognosis, survival, and the expenditure of hospital resources for patients in an intensive-care unit." *New England Journal of Medicine* 305.12 (1981): 667-672.

[5] Ghosh, D., M. Chinnaiyan. "Classification and selection of biomarkers in genomic data using LASSO."*J. Biomed. Biotechnol*, 147-154. 2005.

[6] Gill, Thomas M. "The central role of prognosis in clinical decision making." *Jama* 307.2 (2012): 199-200.

[7] Gottlieb, Scott. "Injury to bile duct during cholecystectomy nearly triples risk of death." BMJ: British Medical Journal 327.7421 (2003): 946.

[8] Guyon, Isabelle, and Andr Elisseeff. "An introduction to variable and feature selection." *The Journal of Machine Learning Research* 3 (2003): 1157-1182.

[9] Kalb, Paul E., and David H. Miller. "Utilization strategies for intensive care units." *Jama* 261.16 (1989): 2389-2395.

[10] Knaus, William A., et al. "APACHE II: a severity of disease classification system." *Critical care medicine* 13.10 (1985): 818-829.

[11] Knaus, William A., et al. "The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults." *Chest Journal* 100.6 (1991): 1619-1636.

[12] R. G. Krishnan, N. Razavian, Y. Choi, S. Nigam, S. Blecker, A. M. Schmidt, and D. Sontag. "Early detection of diabetes from health claims." *Neural Information Processing Systems Workshop: Machine Learning for Clinical Data Analysis and Healthcare*, 2013.

[13] Lee, Joon, David M. Maslove, and Joel A. Dubin. "Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric." Ed. Frank Emmert-Streib. *PLoS ONE* 10.5 (2015): e0127428. PMC. 11 Oct. 2015.

[14] Le Gall, Jean-Roger, Stanley Lemeshow, and Fabienne Saulnier. "A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study." *Jama* 270.24 (1993): 2957-2963.

[15] Metnitz, Barbara, et al. "Austrian validation and customization of the SAPS 3 Admission Score." *Intensive care medicine* 35.4 (2009): 616-622.

[16] Moreno, Rui P., et al. "SAPS 3From 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* 31.10 (2005): 1345-1355.

[17] Nassar, Antonio Paulo, et al. "Caution when using prognostic models: a prospective comparison of 3 recent prognostic models." *Journal of critical care* 27.4 (2012): 423-e1.

[18] Neuvirth, H, M. Ozery-Flato, J. Hu, J. Laserson, M. S. Kohn, S. Ebadollahi, and M. Rosen-Zvi. "Toward personalized care management of patients at risk: the diabetes case study." *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, 395-403. ACM, 2011.

[19] Ng, A. "Feature selection, L1 vs. L2 regularization, and rotational invariance." *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

[20] Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. "Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study." *Lancet Respir Med.* 3(1):42-52. January 2015.

[21] Poole, Daniele, et al. "Comparison between SAPS II and SAPS 3 in predicting hospital mortality in a cohort of 103 Italian ICUs. Is new always better?." *Intensive care medicine* 38.8 (2012): 1280-1288.

[22] Power, G. Sarah, and David A. Harrison. Intensive Care National Audit & Research Centre (ICNARC), London, United Kingdom. "Why try to predict ICU outcomes?" *Current Opinion in Critical Care*. 20(5):544-549, October 2014.

[23] Saeed, Mohammed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. "Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database." 2011; 39:952-960. DOI: 10.1097/CCM.0b013e31820a92c6.

[24] Selker, Harry P., Joni R. Beshansky, and John L. Griffith. "Use of the electrocardiograph-based thrombolytic predictive instrument to assist thrombolytic and reperfusion therapy for acute myocardial infarction: a multicenter, randomized, controlled, clinical effectiveness trial." *Annals of internal medicine* 137.2 (2002): 87-95.

[25] Seneff, Michael, and William A. Knaus. "Predicting patient outcome from intensive care: a guide to APACHE, MPM, SAPS, PRISM, and other prognostic scoring systems." *Journal of Intensive Care Medicine* 5.1 (1990): 33-52.

[26] Silva, Ivanovitch, et al. "Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012." *Computing in Cardiology (CinC)*, 2012. IEEE, 2012.

[27] Silva, Ivanovitch, et al. "Predicting in-hospital mortality of ICU patients: The physioNet/computing in cardiology challenge 2012." *Computing in Cardiology (CinC)*, 2012. IEEE, 2012.

[28] Sinuff, Tasnim, et al. "Mortality predictions in the intensive care unit: Comparing physicians with scoring systems*." *Critical care medicine* 34.3 (2006): 878-885.

[29] Sun, Jimeng, et al. "Combining knowledge and data driven insights for identifying risk factors using electronic health records." *AMIA*. Vol. 2012. 2012.

ACKNOWLEDGMENT

Thank you to Yun Liu for guidance throughout the project, Dr. Collin Stultz for adivising, and Jen Gong for MIMIC data extraction.