Learning Clinical Events

ICU Event Prediction with Multimodal Clinical Data Using Deep Recurrent Attention Networks

Marzyeh Ghassemi*
CSAIL, MIT
Cambridge, MA
mghassem@mit.edu

Harini Suresh*
CSAIL, MIT
Cambridge, MA
hsuresh@mit.edu

Nathan Hunt*
CSAIL, MIT
Cambridge, MA
nhunt@mit.edu

Tristan Naumann
CSAIL, MIT
Cambridge, MA
tjn@mit.edu

Leo Anthony Celi
CSAIL, MIT
Cambridge, MA
lceli@mit.edu

Peter Szolovits
CSAIL, MIT
Cambridge, MA
psz@mit.edu

ABSTRACT

Continuous prediction of clinical outcomes and evaluation of patient risk remains a challenge within intensive care units (ICUs). These learning tasks are complicated by data size, class imbalance, and noisy, heterogeneous data sources.

In this paper, we use a recurrent neural network to integrate the heterogeneous data sources and address the task of continuous risk prediction. We integrate data from static demographic information, free text clinical narratives, vital signs, and lab values. We evaluate these models on four distinct tasks that cover mortality during admission and mortality following discharge. In all cases, prediction is done on a continuous, hourly basis and in a forward-facing manner to approximate "real-time" performance, allowing for the evaluation of RNN models that could inform treatment strategies at the point of care. Further, we consider feature-level attention mechanisms dedicated to generating interpretable network rules, which is necessary for the adoption of such models in practice.

We show that RNNs effectively integrate different data types into a single common representation for a variety of prediction tasks. In addition, we demonstrate that RNNs are able to provide an early warning for each task, often with enough time for clinically actionable planning.

CCS CONCEPTS

•**Applied computing** → **Health care information systems;** *Consumer health;* •**Computing methodologies** → **Supervised learning by classification;** *Neural networks;*

KEYWORDS

Clinical Prediction; Machine Learning; Data Science

ACM Reference format:
Marzyeh Ghassemi, Harini Suresh*, Nathan Hunt*, Tristan Naumann, Leo Anthony Celi, and Peter Szolovits. 2017. Learning Clinical Events. In *Proceedings of Knowledge, Discovery, and Data Mining Conference, Halifax, Nova Scotia, August 16 - 19, 2017 (KDD 2017),* 9 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

*The first three authors contributed equally to this work.

KDD 2017, Halifax, Nova Scotia

DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Intensive Care Units (ICUs) play an increasing role in acute healthcare delivery [30], and clinicians must make quick and accurate decisions about patient care. Clinical decision-making is often made in settings of limited knowledge and high uncertainty; for example, only 10 of the 72 ICU interventions evaluated in randomized controlled trials (RCTs) are not associated with improved outcomes [27].

Our goal is to gain insight from healthcare data that has already been collected for the primary purpose of facilitating patient care. The secondary analysis of healthcare data is a critical step toward improving modern healthcare, as it affords the study of care in the real care settings and patient populations. Existing RCTs do not cover a majority of treatments that are commonly used [24, 25], and even those that are commonly used contain structural biases in subject recruitment [29].

Electronic Health Record (EHR) systems that meet federal requirements are present in most acute care hospitals (97% in 2014 [4]) and office-based physicians (78% in 2015 [26]). This availability allows new investigations into evidence-based decision support, where we can learn when patients are at high risk for mortality or need a given intervention. Unlike traditional measures of risk and treatment, which are often evaluated at single endpoints (e.g., in-hospital mortality), we seek models that account for evolving clinical information throughout the patient's stay.

The task of such continuous "forward-facing" event prediction on a binned basis (e.g., every t hours) is particularly applicable in the ICU setting. Similar efforts have previously considered using a topic representation of clinical notes to predict mortality [10], or Cox proportional hazard models to forecast the time to septic shock onset [15]. Others have used representations of patient physiological signals, including multi-task Gaussian processes, to predict mortality [11], or switching-state autoregressive models to forecast the onset of interventions [12]. Most recently, recurrent neural networks (RNN) have been applied in modeling sequential EHR data to tag ICU signals with billing code labels [6, 23] and to identify the impact of different drugs for diabetes [21]. New work has also

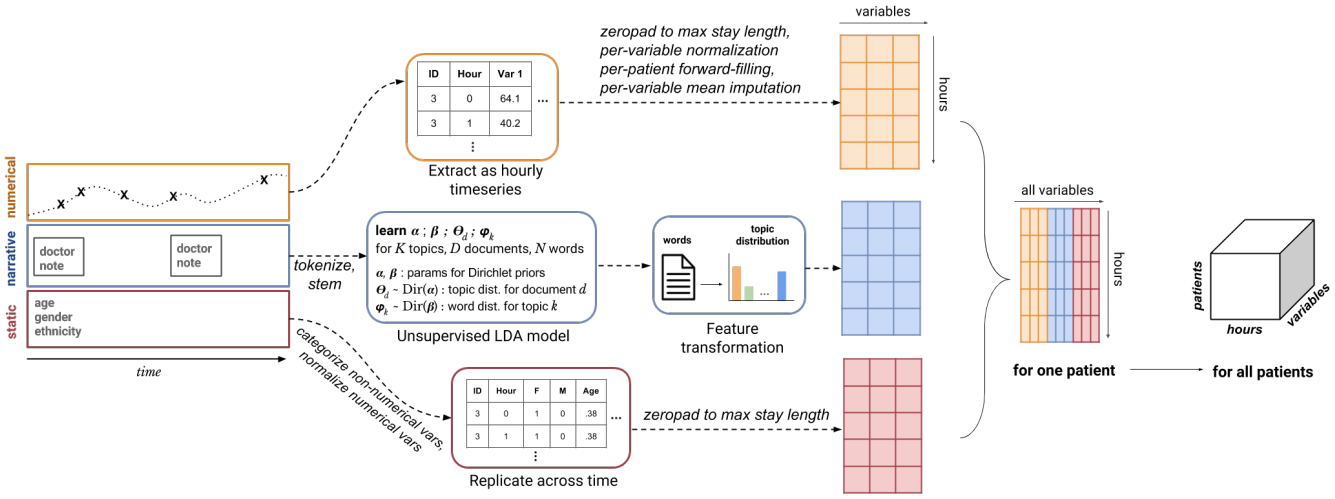


Figure 1: A visual representation of the data used. 1) *Numerical data*, including vitals and lab tests. The timestamp for each data point is rounded to the nearest hour, and hours with multiple measurements for a variable are assigned the average of those measurements. Each measurement is normalized according to the min and max for that var and each patient's data are zero-padded to the maximum stay length (240 hours). To fill in missing values, we forward-fill values for each patient, and mean-impute for any remaining missing values. 2) *Narrative data*, which consists of unstructured text notes. After preprocessing, LDA is used to obtain underlying topics and we then represent each note as a distribution over these topics. We forward-fill and aggregate these topic vectors across time, mean-imputing any values that are still missing. 3) *Static Data*, including variables recorded at admission such as sex, age, and ethnicity. Categorical variables such as ethnicity and ICU type are transformed into one-hot vectors containing each possible type. We replicate this data across time so that we are able to feed in this information at every timestep. We normalize numerical values and use forward-filling and imputation as before.

introduced the use of temporal attention for use in early diagnostic prediction of chronic diseases from time-ordered billing codes [7].

In this work, we use data from the publicly-available Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III) database [18] to address early ICU event prediction in four distinct clinical tasks that span ICU, in-hospital, and post-discharge mortality. In all cases, we use representations of ICU patient data from each available data type[1] — vitals (~hourly), labs (~daily), demographics (static after observation at admission), and notes (typically every 12 hours) — to model target outcomes in varying-length patient records. We use RNNs to model variable-length timeseries [1, 17] and consider the task of early forward-facing prediction. This is an important problem in the ICU because each patient's severity of illness is constantly evolving, and clinicians can use dynamically computed risk scores to schedule staff, make bed allocations, or decide who to should be discharged to the floor.

Specifically, we consider the following four predictive tasks: ICU mortality, hospital mortality, and both 30 and 90 day post-discharge mortality. We contextualize our results with respect to prior work and provide possible interpretations of the attention-level features learned by our models. In doing so, we make the following contributions:

- We provide a method that integrates all data modalities — vitals, labs, demographic, and notes — toward the prediction of mortality in MIMIC-III, as opposed to existing literature leveraging only a subset of these data.
- We perform forward-facing, hourly prediction of clinical risk factors that could be used at the time of care. To maintain this use case, we do not leverage information that are often recorded at discharge (e.g., billing codes, discharge summaries).
- We provide a single RNN to model outcomes given any variable-length patient record, rather than segmenting patient data and training separate models for each segment.

2 METHODS

In this section, we detail the data and preprocessing (Section 2.1), the learning tasks (Section 2.2) and provide an overview of RNNs (Section 2.3) as well as the experimental settings (Section 2.4).

2.1 Data and Preprocessing

Figure 1 provides an overview of our data extraction process. For every patient, we extracted 1) static clinical features, including age, sex, and SAPS-II score; 2) the de-identified clinical notes; 3) the nurse-verified vitals; and 4) the reported lab values. We represent notes via an LDA topic model [10], forward-fill vitals and labs data [6], and concatenate static data to the evolving information [9].

[1]MIMIC-III also contains high-frequency waveforms, but they could not be integrated into our model due as they lack timestamps.

We train RNNs to predict clinical events using combinations of the available data for all tasks.

We use data from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III) Database [18]. MIMIC-III v1.4 contains data from over 58,000 hospital admissions of approximately 38,600 adults from 2001–2012, and is publicly available. We use data from the first ICU stay of patients in the Medical Care Unit (MICU), Cardiac Care Unit (CCU), Cardiac Surgery Recovery Unit (CSRU), Surgical Intensive Care Unit (SICU), and Trauma Surgical Intensive Care Unit (TSICU).

We consider only patients older than 15 who were in the ICU for between 12 and 240 hours. [2] This yields 34,148 unique ICU stays. We consider only each patient's first ICU stay to avoid training and testing on data from the same patient. For each patient, we extract 5 static variables (gender, age, ethnicity, admission type, ICU unit), 29 time-varying vitals and labs (anion gap, bicarbonate, blood urea nitrogen, chloride, creatinine, diastolic blood pressure, fraction inspired oxygen, Glascow coma scale total, glucose, heart rate, hematocrit, hemoglobin, lactate, magnesium, mean blood pressure, oxygen saturation, partial thromboplastin time, phosphate, platelets, potassium, prothrombin time, international normalized ratio of the prothrombin time (INR), respiratory rate, sodium, systolic blood pressure, temperature, weight, white blood cell count, blood pH), and all available notes for each patient as timeseries across their entire stay. We remove any data that occurs within 24 hours of patient death or discharge as these data are very close to clinicians predicting a patient outcome, and have little value in a real-time prediction task.

Vital and lab measurements are given timestamps that are rounded to the nearest hour, and if an hour has multiple measurements for a signal those measurements are averaged. Each variable is normalized using the minimum and maximum of that variable across all patients. Since there are many missing values, we then forward-fill missing values for each patient. Any remaining missing values are imputed with the population mean for that variable.

For clinical narrative notes, Latent Dirichlet Allocation [2, 14] is used to generate underlying topics, and the notes are then represented as a distribution over these topics. We use the settings that achieved the best performance in other work [10], namely 50 topics that result in a 50-dimensional vector of topic proportions for each note. Since notes occur less frequently than every hour, we replicate forwards and aggregate the note vectors through time. For example, if a patient had a note A recorded at hour 3 and a note B at hour 7, hours 3–6 would contain the topic distribution from A, while hours 7 onward would contain the aggregated topic distribution from A and B combined. Dataset statistics are shown in Table 1, and the top 10 most likely words for each learned topic are presented in Table 2.

Static variables were replicated across all timesteps for each patient. Categorical variables such as sex, ethnicity, or ICU type were transformed into their own binary one-hot vectors. After the transformation, we end up with 52 total features representing the static variables. Numerical variables such as age were forward-filled as before.

[2]Young patients are excluded as they typically exhibit different physiology from an adult population.

The physiological variables, topic distribution, and static variables for each patient are concatenated into a single feature vector per patient per hour.

2.2 Task Definitions

We evaluate our model on the prediction of four tasks: 1) ICU mortality (ICU-Mort), 2) hospital mortality (Hosp-Mort), and both 3) 30 day post-discharge mortality (30-Mort) and 4) 90 day post-discharge mortality (90-Mort).

We consider these four prediction tasks for their clinical relevance and in order to draw comparisons with existing work.

ICU-Mort and Hosp-Mort present a risk-score that anticipates the risk of mortality, and can be used as proxies for severity of illness. These proxies are useful for risk stratification [28], resource utilization [19] and clinical decision-making [13]. Additionally, 30-Mort and 90-Mort tasks allow clinicians to make more informed decisions about important and specific treatment paths that are well-established following onset.

2.3 RNN Models

We use long short-term memory networks (LSTM) [17], a variant of RNNs. LSTMs are used because of their ability to effectively model varying-length timeseries data and capture long-term dependencies [1]. They are well-suited to modeling clinical data because evidence of certain conditions may be spread apart over several hours or days, and important symptoms may present early on in a patientfis stay.

We also test the ability to predict mortality with an attention mechanism. Neural networks have used attention mechanisms to visualize the factors that are most important in generating image descriptions [32], answer questions from text [16], and process speech [8]. When producing outputs, the attention mechanism typically allows the network to refer to specific states from past timesteps, instead of a single fixed-length representation of the state at the current timestep [23]. In most implementations, attention can be thought of as a weighted combination of all internal memory locations, rather than a single discrete location.

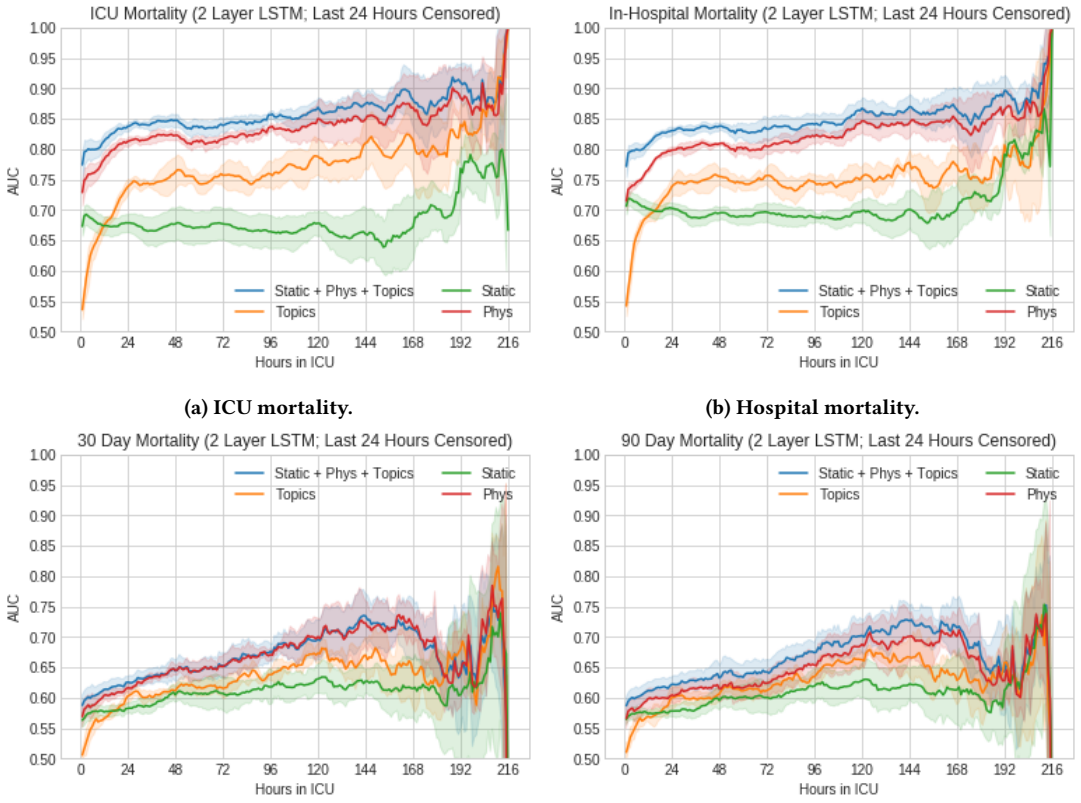
Previous work usually uses attention over past timesteps. In our case, we attend over the variables within a single timestep, since we care specifically about the variables that are most important, rather than the just the timesteps. We call this method *fine-grained attention*. We calculate weights for each variable in a timestep as a function of the variables themselves and the previous hidden state of the LSTM. The input variables are then multiplied by these weights, giving a weighted input vector that is fed into the LSTM layers. Furthermore, we can use these weights to visualize what was important in the input right before we predicted the onset of a given intervention.

At a given timestep t, having seen the input sequence $x_1 \dots x_{t-1}$ for a given patient, and given the current input x_t, we predict $\hat{y_t}$, the probability of the target outcome y_t:

$$z_t = ATTN(x_t) \tag{1}$$

$$h_t = LSTM(z_t) \tag{2}$$

$$\hat{y_t} = W_y h_t + b_y \tag{3}$$



(a) ICU mortality.

(b) Hospital mortality.

(c) 30-day post-discharge mortality.

(d) 90-day post-discharge mortality.

Figure 2: RNN model performance every hour after ICU admission measured via AUC on four clinical tasks: a) ICU mortality, b) in-hospital mortality, c) 30 day post-discharge mortality, and d) 90 day post-discharge mortality. In each case, the data and models are as described in Sections 2.2 and 2.3. Our prediction tasks vary in complexity, but all lose train/test set support as time goes on because fewer patients have long ICU stays. For example, as time goes on it becomes more likely ICU patients will need ventilation. In this task, the number of test set control examples still in the ICU becomes smaller than the number of positives around 24 hours, which is when the task begins to become progressively more difficult.

where $z_t, x_t \in \mathbb{R}^V, W_y, h_t \in \mathbb{R}^{L_2}, b_y \in \mathbb{R}$ where L_1, L_2 are the first and second hidden layer sizes, respectively, and V is the dimensionality of the input (number of variables). When we do not use attention, x_t is simply the input to the LSTM instead of z_t.

ATTN operates by taking a function of the input and previous hidden state, and create a new "weighted" input z_t:

$$a_t = W_h h_{t-1} + W_x x_t + b_a \tag{4}$$

$$\beta_{x_t} = \text{softmax}(\tanh(v \odot a_t)) \tag{5}$$

$$z_t = \beta_{x_t} \odot x_t \tag{6}$$

where $W_h \in \mathbb{R}^{V \times L_1}$, $W_x \in \mathbb{R}^{V \times V}$, and $v, b_a \in \mathbb{R}^V$ are learned parameters, and $a_t, z_t, \beta_{x_t} \in \mathbb{R}^V$.

LSTM performs the following update equations for a single layer, given its previous hidden state and the new input:

$$f_t = \sigma(W_f[h_{t-1}, z_t] + b_f) \tag{7}$$

$$i_t = \sigma(W_i[h_{t-1}, z_t] + b_i) \tag{8}$$

$$\tilde{c_t} = \tanh(W_c[h_{t-1}, z_t] + b_c) \tag{9}$$

$$c_t = f \odot c_{t-1} + i \odot \tilde{c_t} \tag{10}$$

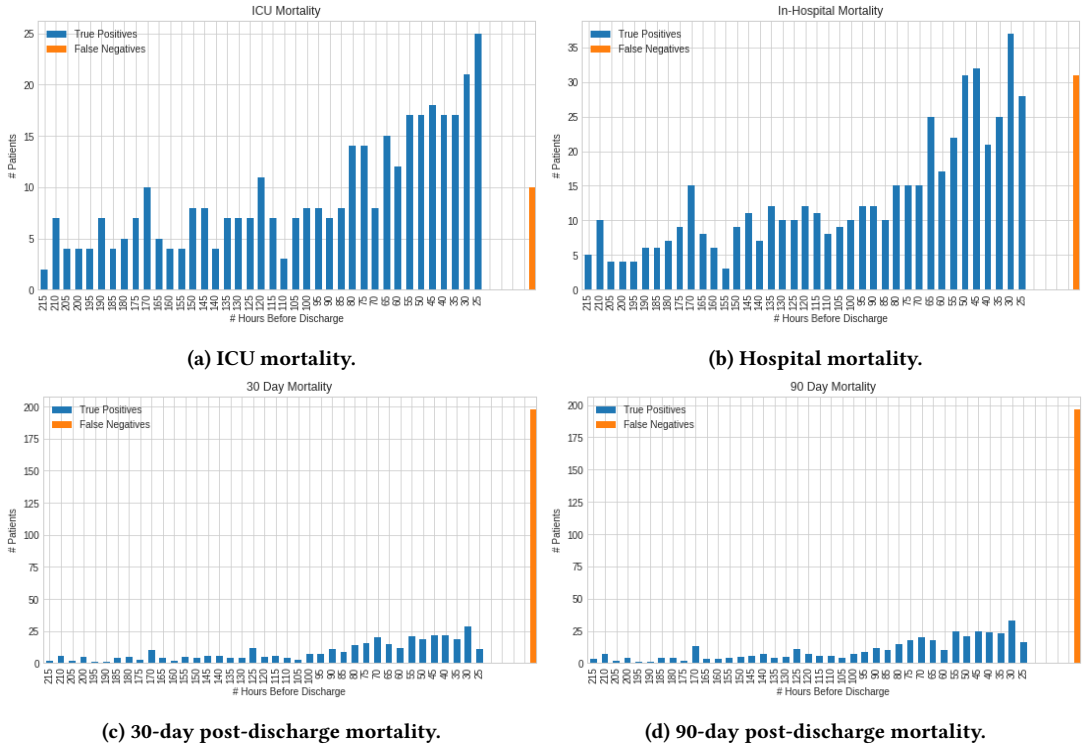
$$o_t = \sigma(W_o[h_{t-1}, z_t] + b_o) \tag{11}$$

$$h_t = o_t \odot tanh(c_t) \tag{12}$$

where $W_f, W_i, W_c, W_o \in \mathbb{R}^{L_1 \times (V+L_1)}$, $b_f, b_i, b_c, b_o \in \mathbb{R}^{L_1}$ are learned parameters, and $f_t, i_t, \tilde{c_t}, c_t, o_t, h_t \in \mathbb{R}^{L_1}$.

As before, when we do not use attention, x_t is used as input rather than z_t. This is generalized to multiple layers by providing h_t from the previous layer in place of the input.

In these equations, σ stands for an element-wise application of the sigmoid (logistic) function, and $\odot$ is an element-wise product.



(a) ICU mortality.

(b) Hospital mortality.

(c) 30-day post-discharge mortality.

(d) 90-day post-discharge mortality.

Figure 3: Early warning value of RNN predictions on four clinical tasks: a) ICU mortality, b) in-hospital mortality, c) 30 day post-discharge mortality, and d) 90 day post-discharge mortality. All tasks are censored at 24 hours prior to ICU discharge; therefore, we show all patients that were not correctly predicted by this gap in orange. The "earliest" time we can predict a patient correctly is immediately is 215 hours before onset (e.g., after observing 4 hours of data in a record where onset occurs at 220). The "latest" time that we can correctly predict the event is 25 hours prior to onset, and patients from this time are shown separately.

We calculate classification loss using binary cross-entropy. At each timestep t, the loss for predictions for N patients is:

$$\mathcal{L}(\hat{y}_t^1 \dots \hat{y}_t^N) = -\frac{1}{N}\sum_{i=1}^{N} y_t^i \log \hat{y}_t^i \tag{13}$$

To get the total loss for a set of examples, we just sum the losses at each timestep.

2.4 Experimental Settings

We assemble an evaluation dataset from our full cohort of data for N patients. The data was split into training/validation/testing sets with a 70/10/20 split, and stratified on in-hospital mortality in order to have a spectrum of patient severity in both the train and test sets.

We investigated several combinations of layer sizes and chose to use two layers of size 1024 and 256 hidden nodes for reported results based on cross-validation results. We implemented all models in TensorFlow version 0.12.1 using the Adam optimizer on mini-batches of 400 patients.

For mortality prediction, we include only timesteps until 24 hours before the end of the sequence (discharge or death). Patients with less than 25 hours of ICU data were thus discarded, so that the model always has some data from which to predict. Forcing the model to learn to predict with such a gap, rather than predicting mortality shortly before it occurs, makes the model much more useful in practice; it may be able to inform a physician before they are aware of the acuity of a patient's state.

3 RESULTS

3.1 Forward-facing Prediction of Clinical Events

We evaluated the predictive power of each data type and outcome pair in the RNN models. Figures 2a through 2d show the AUCs of predicting mortality during hospital stay (ICU in Figure 2a and hospital in Figure 2b), and post-discharge mortality (30-day in Figure 2c and 90-day in Figure 2d). In all figures, at each time, we are predicting only for the pool of patients who have not yet had an onset of the targeted task, and will not have one for at least 24 hours.

The different data types contributed differing amounts of information to our clinical tasks. We found that static data (e.g., age, gender, and other demographic information) tended to be least valuable over time consistently. Topics derived from the clinical notes in a patient's record were often next-best, increasing predictive

performance for most tasks by a significant margin. As shown in Table 2, the the top words for the topics learned compare to those in prior work on LDA modeling of clinical notes.[10] Physiological vitals and labs (e.g., heartrate, blood pressure, respiration rate, blood glucose level, etc.) performed similarly to the topics data in many tasks, but were better than topics data in ICU and hospital mortality prediction.

In all cases, combining all data types performed best over time, but this was not always significantly improved over the physiological data on its own. The value of adding topics and static data to the physiological data also tended to decrease as patients were in the ICU for longer stays. Note that the average length of an ICU stay is 3–4 days (72–96 hours). In most tasks, the standard deviation of performance across folds tended to increase after 96 hours, which may be due to the increase in variability of very sick patients in a randomly sampled test sets over the five runs.

Both mortality prediction tasks during hospital admission had consistent AUC performance over the course of nine days considered. Post-discharge mortality tasks were generally harder, with significantly lower AUCs. This is sensible, as we would expect that information over a hospital admission may not be as relevant to post-discharge outcomes when there are other factors present (e.g., patients with chronic health problems).

3.2 Potential Use as Early Warning Score

We consider the task of evaluating the RNN model predictions using all data types to produce an early warning for each clinical task. In this setting, we recompute the risk score for each patient in the test set every hour. We say that a patient has been identified "early" for the clinical event the first time that their computed score crosses a threshold computed over the validation set. This threshold varies over time for each clinical event, but was chosen for each task and time to balance specificity and sensitivity. This corresponds to the point closest to (0,1) on the ROC curve.

Figures 3a through 3d show our early prediction time for each of the four clinical tasks. As noted, the "earliest" early prediction time possible is 215 hours before onset, and the "latest" time is 25 hours before onset. All other patients are considered unidentified within an appropriate amount of warning time.

Early patient identification varied based on the task complexity. We predict first intervention onset early in most of the patients. For ICU mortality, we predict 362/372 patients early with a mean early warning time of 92 hours. For in-hospital mortality, we predict 513/544 patients early at a mean of 88 hours before ICU discharge/death. Predicting mortality after the patients leave the hospital is much more difficult. We still identify a majority of patients early (358/556 patients who will die within 30 days and 398/595 who will die within 90 days), but the number of patients we correctly identify decreases. Adding patients to the training set who are healthier in the present and will only die later seems to reduce the models ability to identify patients who will be in an acute state in the nearer future. It may thus be advisable to train separate models for near-term and long-term mortality prediction or patient acuity scoring.

3.3 Interpretation of Feature-Level Weights for Tasks

Using the model with attention over input variables, we can visualize the distribution of of feature weights when it predicts a given outcome. A distribution of weights that is concentrated around a higher value may be an indication that the variable is always important to the prediction. Bimodal distributions may be a sign of the complex interactions within variables: a specific variable may be weighted highly depending on the other variables present and the sequence of inputs the LSTM has seen previous to this point. A distribution with a large standard deviation likely indicates that the variable doesn't play a consistent role in prediction of the task.

For the task of ICU-Mort, for example, variables such as temperature display a bimodal distribution (Figure 4a) while pH (Figure 4b) has a broader distribution with a higher mean. In other words, dependent on other factors, temperature's importance varies between two modes, while the importance of pH is on average higher but more uniformly distributed. Fraction inspired oxygen has a very narrow distribution centered on a single value, suggesting that it consistently has a relatively low importance in predicting ICU mortality (Figure 4c).

4 RELATED WORK

Current ICU practice evaluates patient acuity using scoring systems based on static periods of patient data like SAPS-II [22], SOFA [31], or APACHE [20]. Such scores are also evaluated at a single end point, such as in-hosptial mortality or mortality 28 days post-discharge. Single risk scores are unable to capture the different ways in which a patient may be ill.

Clinical tasks have been considered in prior work on the current MIMIC-III dataset, or the previous MIMIC-II dataset.

For mortality prediction during hospital stays, we achieved AUCs of 0.83/0.84 at the 24 and 48 hour prediction marks for in-hospital mortality, which is comparable to previous work. Ghassemi et al. used topics derived from the MIMIC-III notes on a forward-facing prediction task to achieve an AUC of 0.84/0.85 at 12 and 24 hours for in-hospital mortality [10]. However, they trained a separated model for each 12 hour task and did not integrate the physiological vitals data. Caballero et al. employed a Latent Dynamic System model for prediction of patient mortality using notes and medication to achieve an AUC of 0.86 on 24 hour in-hospital mortality [3]. Che et al. used an LSTM model on the vitals data to obtain an AUC of 0.82 on 48 hour mortality prediction [5], and this was improved to 0.85 when using a GRU network that included a model of missing data on vitals and extracted medications [6].

Our performance on 30-day post-discharge mortality at 24 and 48 hours was lower than previously reported work on the MIMIC-II dataset (0.64 and 0.66 compared to 0.76/0.78), which may be due to the transition of the MIMIC-III database system to a new backend that reports information differently.[3]

[3]The original Philips CareVue system archived data in MIMIC-II over the period from 2001–2008. This was replaced in 2008 with the Metavision data management system, which was added to the MIMIC-III dataset.

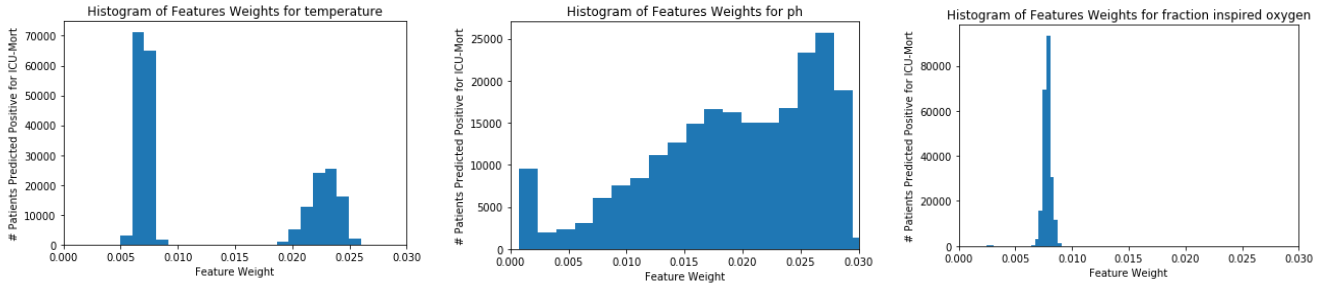


Figure 4: a) Histogram of weights for the temperature variable in the timestep when we predict ICU mortality for a patient. This distribution is bimodal, indicating that the importance of temperature adopts one of two modes depending on the other variables present and already observed. b) Histogram of weights for the pH variable in the timestep when we predict ICU mortality for a patient. This distribution is broadly distributed but concentrated at higher values than other variables. c) Weights for fraction inspired oxygen have one narrow peak, indicating that it consistently has a relatively low importance in the prediction task.

5 CONCLUSIONS

Hospitals are highly uncertain environments where clinical staff must make decisions about patient care in real-time with noisy heterogeneous data. Current clinical evidence is based on expensive and rare RCTs that do not cover many common treatments, and the increasing prevalence of EHRs offers new opportunities to create evidence-based decision support.

In this work, we demonstrated that real ICU data can be used in the forward-facing prediction of several important clinical tasks. Rather than focusing on a supervised model for a single clinical prediction, our work focuses on learning models that generalize across time and tasks. We depart from prior work by investigating the relative value of each type of available data towards our tasks, and evaluating the amount of early warning time that such models provide. We also provide a methodology for investigating important features for a particular prediction task using a feature-level attention mechanism. This creates a single recurrent model that allows for variable-length patient data to be evaluated for a robust set of clinical risks.

The models explored in this work, and the results we have obtained, are a first step towards our ultimate goal of clinically actionable predictions that could be used to predict important clinical events in a real hospital setting.

ACKNOWLEDGMENTS

This research was funded in part by the Intel Science and Technology Center for Big Data and the National Library of Medicine Biomedical Informatics Research Training grant (NIH/NLM 2T15 LM007092-22). The authors would like to thank Raziyeh Elise, Abbas Benjamin, and Unnamed Marit for their feedback and support during this research.

REFERENCES

[1] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5, 2 (1994), 157–166.

[2] D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet allocation. *JMLR* 3, 5 (2003), 993–1022.

[3] Karla L Caballero Barajas and Ram Akella. 2015. 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*. ACM, 69–78.

[4] Dustin Charles, Meghan Gabriel, and Michael F Furukawa. 2013. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2012. *ONC data brief* 9 (2013), 1–9.

[5] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. 2015. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 507–516.

[6] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2016. Recurrent neural networks for multivariate time series with missing values. *arXiv preprint arXiv:1606.01865* (2016).

[7] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *Advances in Neural Information Processing Systems*. 3504–3512.

[8] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*. 577–585.

[9] Cristóbal Esteban, Oliver Staeck, Stephan Baier, Yinchong Yang, and Volker Tresp. 2016. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*. IEEE, 93–101.

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

[11] Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. 2015. A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data. In *Proc. Twenty-Ninth AAAI Conf. on Artificial Intelligence*.

[12] Marzyeh Ghassemi, Mike Wu, Mengling Feng, Leo A Celi, Peter Szolovits, and Finale Doshi-Velez. 2016. Understanding vasopressor intervention and weaning: Risk prediction in a public heterogeneous clinical time series database. *Journal of the American Medical Informatics Association* (2016), ocw138.

[13] Thomas M Gill. 2012. The central role of prognosis in clinical decision making. *Jama* 307, 2 (2012), 199–200.

[14] T. Griffiths and M. Steyvers. 2004. Finding Scientific Topics. In *PNAS*, Vol. 101. 5228–5235.

[15] Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. 2015. A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine* 7, 299 (2015), 299ra122–299ra122.

[16] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*. 1693–1701.

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[18] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3 (2016).

[19] Paul E Kalb and David H Miller. 1989. Utilization strategies for intensive care units. *Jama* 261, 16 (1989), 2389–2395.

[20] William A Knaus, DP Wagner, EA e a1 Draper, JE Zimmerman, Marilyn Bergner, P Gl Bastos, CA Sirio, DJ Murphy, T Lotring, and A Damiano. 1991. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *CHEST Journal* 100, 6 (1991), 1619–1636.

[21] Rahul G Krishnan, Uri Shalit, and David Sontag. 2015. Deep kalman filters. *arXiv preprint arXiv:1511.05121* (2015).

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

[23] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. 2015. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677* (2015).

[24] J Michael McGinnis, Leigh Stuckhardt, Robert Saunders, Mark Smith, and others. 2013. *Best care at lower cost: the path to continuously learning health care in America.* National Academies Press.

[25] Edward J Mills, Kristian Thorlund, and John PA Ioannidis. 2013. Demystifying trial networks and network meta-analysis. *Bmj* 346 (2013), f2914.

[26] Office of the National Coordinator for Health Information Technology. 2016. Office-based Physician Electronic Health Record Adoption. *Health IT Quick-Stat 50* (2016).

[27] Gustavo A Ospina-Tascón, Gustavo Luiz Büchele, and Jean-Louis Vincent. 2008. Multicenter, randomized, controlled trials evaluating mortality in intensive care: Doomed to fail? *Critical care medicine* 36, 4 (2008), 1311–1322.

[28] Michael Seneff and William A Knaus. 1990. 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.

[29] Justin Travers, Suzanne Marsh, Mathew Williams, Mark Weatherall, Brent Caldwell, Philippa Shirtcliffe, Sarah Aldington, and Richard Beasley. 2007. External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? *Thorax* 62, 3 (2007), 219–223.

[30] Jean-Louis Vincent. 2013. Critical care-where have we been and where are we going? *Critical Care* 17, Suppl 1 (2013), S2.

[31] J-L Vincent, Rui Moreno, Jukka Takala, S Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PM Suter, and LG Thijs. 1996. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive care medicine* 22, 7 (1996), 707–710.

[32] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.. In *ICML*, Vol. 14. 77–81.

A DATASET STATISTICS

Table 1: Dataset Statistics

	Train	Test	Total
Patients	27,318	6,830	34,148
Notes	564,652	140,089	703,877
Elective Admission	4,536	1,158	5,694
Urgent Admission	746	188	934
Emergency Admission	22,036	5,484	27,520
Mean Age	63.9	64.1	63.9
Black/African American	1,921	512	2,433
Hispanic/Latino	702	166	868
White	19,424	4,786	24,210
CCU (coronary care unit)	4,156	993	5,149
CSRU (cardiac surgery recovery)	5,625	1,408	7,033
MICU (medical ICU)	9,580	2,494	12,074
SICU (surgical ICU)	4,384	1,074	5,458
TSICU (trauma SICU)	3,573	861	4,434
Female	11,918	2,924	14,842
Male	15,400	3,906	19,306
ICU Mortalities	1,741	439	2,180
In-hospital Mortalities	2,569	642	3,211
30 Day Mortalities	2,605	656	3,216
90 Day Mortalities	2,835	722	3,557
Vasopressor Usage	8,347	2,069	10,416
Ventilator Usage	11,096	2,732	13,828

B EXTRACTED TOPICS

Table 2: Word To Topic Mappings

Topic	Top Words
1	plan assessment response action afib failure gtt acute atrial hr
2	contrast ct right clip left pelvis chest reason small iv
3	ml dl mg pm meq icu code medications continue total
4	gtt hr pt propofol neo wean min given sedated neuro
5	severe echo pulmonary systolic cardiac patient ef left aortic heart
6	pt neuro plan status head response assessment ct commands action
7	pt clear neuro hr resp gi gu abd pain urine
8	bowel abdomen abdominal ct small air pelvis free obstruction contrast
9	mg ml continue patient po daily cardiac history pm pt
10	history patient po daily pain mg icu past given ed
11	pt impaired activity sit status mobility balance stand functional supine
12	chest reason clip ap portable left right old year examination
13	hct bleed bleeding gi blood stable units gib pt prbc
14	ml pm mg dl continue meq respiratory rr min hour
15	tracing sinus previous rhythm wave st atrial compared ventricular left
16	fracture trauma fx left right fractures fall rib ct hip
17	ml mg dl pm meq icu total medications extremities rhythm
18	procedure stitle dr catheter patient placement picc line drain placed
19	etoh withdrawal abuse ciwa alcohol pt valium ativan seizure seizures
20	pericardial cath effusion stemi cardiac ccu lab tamponade echo drain
21	hd pt hypotension ed bp sepsis vanco plan given renal
22	pain pt plan response assessment action given monitor hr control
23	tube chest placement reason right clip line tip left ap
24	stroke left cva ct weakness right head heparin sided neuro
25	lymphoma fever infection patient abscess tpn picc bmt pain fevers
26	cancer lung pleural ct mass metastatic ca chest effusion right
27	pt pain plan assessment sats response continue patient cough action
28	ml pm mg dl min meq icu pulse present mmhg
29	date order ml mg pain present iv tube normal absent
30	contrast hemorrhage head ct right left clip reason old year
31	ml assessed dl mg pulse pm meq icu right left
32	dl mg weight meq nutrition diet pm body arterial patient
33	response action assessment pt plan failure acute continue monitor hr
34	pain pacer assessment plan response wires cabg pacemaker temporary
35	tube aspiration trach intubation airway respiratory chest lung intubate
36	spine fracture cervical clip reason contrast ct spinal soft report
37	insulin gtt blood addendum dm section protected bs scale diabetes
38	valve normal left aortic mitral ventricular leaflets right mildly mild
39	left right dvt femoral lower extremity reason clip old vein
40	pt vent care resp secretions intubated remains respiratory abg plan
41	artery right carotid left numeric identifier clip aneurysm internal contr
42	pt family patient ni care home time support daughter wife
43	pt hr lasix po ? given sats bp resp oob
44	pt hr ni today noted remains bp iv note micu
45	skin wound area pt care applied dressing plan coccyx continue
46	sounds lung assessment ventilation breathing comments airway type i
47	likely ml renal given pending status mg urine negative ct
48	pt arrest cath cardiac transferred cad cabg ccu admitted osh
49	meq mg valuables dl transferred pmh rate bp heart date
50	liver hepatic renal transplant portal reason right cirrhosis biliary ascite