Multitask Learning and Benchmarking with Clinical Time Series Data

Hrayr Harutyunyan, Hrant Khachatrian
YerevaNN; Yerevan State University
Yerevan, Armenia
{hrayr,hrant}@yerevann.com

David C. Kale, Aram Galstyan
USC Information Sciences Institute
Marina del Rey, CA 90292
{kale,galstyan}@isi.edu

ABSTRACT

Health care is one of the most exciting frontiers in data mining and machine learning. Successful adoption of electronic health records (EHRs) created an explosion in digital clinical data available for analysis, but progress in machine learning for healthcare research has been difficult to measure because of the absence of publicly available benchmark data sets. To address this problem, we propose four clinical prediction benchmarks using data derived from the publicly available Medical Information Mart for Intensive Care (MIMIC-III) database. These tasks cover a range of clinical problems including modeling risk of mortality, forecasting length of stay, detecting physiologic decline, and phenotype classification. We formulate a heterogeneous multitask problem where the goal is to jointly learn multiple clinically relevant prediction tasks based on the same time series data. To address this problem, we propose a novel recurrent neural network (RNN) architecture that leverages the correlations between the various tasks to learn a better predictive model. We validate the proposed neural architecture on this benchmark, and demonstrate that it outperforms strong baselines, including single task RNNs.

ACM Reference format:
Hrayr Harutyunyan, Hrant Khachatrian and David C. Kale, Aram Galstyan. 2017. Multitask Learning and Benchmarking with
Clinical Time Series Data. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 11 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In the United States alone, each year over 30 million patients visit hospitals [17], 83% of which now use a basic electronic health record (EHR) system [26]. This trove of digital clinical data represents a significant opportunity for researchers in data mining and machine learning to solve pressing health care problems such as early detection and triage of at-risk patients and identification of high-cost patients [6]. Such problems are not new,[1] but the increased availability of digital clinical data and recent high-profile machine learning successes [20, 57] have sparked a renewed interest in them.

While there has been a steady growth in machine learning research for health care,[2] several substantial obstacles have slowed our progress in fully harnessing the value of digital health data. One of the largest is the absence of universally accepted benchmarks to evaluate competing modeling approaches. Such benchmarks play an important role in accelerating progress in machine learning research by focusing the community and facilitating reproducible research and competition. The winning error rates in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) plummeted almost an order of magnitude from 2010 (0.2819) to 2016 (0.02991) [30]. In contrast, practical progress in machine learning for health care has been difficult to measure because of the use of different cohorts, task definitions, and input variables [9, 23, 36, 40].

In this paper we propose public benchmarks for four different clinical problems: mortality prediction, detection of physiologic decompensation, forecasting length of stay (LOS), and phenotyping. Our benchmarks share a common data set derived from the Medical Information Mart for Intensive Care (MIMIC-III) database [32]. These data are suitable for research on a variety of machine learning topics, such as learning with skewed target distributions and time series analysis. The code to build our benchmarks is available online at https://github.com/yerevann/mimic3-benchmarks, and we will continue to curate and expand the benchmarks over time.

In addition to establishing results for a number of strong baselines, we also describe a multitask recurrent neural network architecture that learns to solve all four problems jointly, capturing potential long-term dependencies between tasks. While heterogeneous multitask learning has been studied (and in fact was first proposed) in the context of clinical prediction by Caruana et al. [10], our setting differs because our targets are heterogeneous in both output type and temporal structure, e.g., length of stay (regression) is predicted continuously, while in-hospital mortality (classification) is predicted once early in admission. Our proposed RNN architecture includes one or more long short-term memory (LSTM) layers shared across tasks, followed by task-specific multilayer perceptron (MLP) outputs that are temporally aligned to their respective tasks. One of the benefits of this modular architecture is that it can be easily generalized to larger numbers of diverse prediction problems.

We summarize our main technical contributions here:

(1) We use the MIMIC-III database to derive four first-of-their-kind standardized benchmarks for machine learning researchers working on clinical modeling problems.
(2) We formulate a heterogeneous multitask LSTM architecture to jointly model these clinical prediction problems.

[1] The word *triage* dates back to at least World War I and possibly earlier [31], while the Apgar risk score was first published in 1952 [5].

[2] Inspiring a conference by that name, the *Machine Learning for Healthcare Conference.*

Conference'17, Washington, DC, USA

DOI: 10.1145/nnnnnnn.nnnnnnn

(3) We propose a custom loss function that consists of a weighted combination of the individual task losses. This loss function improves overall performance by ensuring that the model makes reasonable progress on all learning tasks.
(4) We demonstrate in experiments that the proposed multi-task LSTM outperforms single-task baselines on three out of our four benchmark problems and is competitive on the fourth.

The remainder of the paper is organized as follows: We define our benchmark tasks in Section 2 and then provide a comprehensive survey of related work in Section 3. Section 4 describes our proposed LSTM architecture. Section 5 describes our benchmark data set. In Section 6 we present our experimental results and discuss their implications. We conclude the paper by identifying some open problems and future research directions in Section 7.

2 BENCHMARK PREDICTION TASKS

The tasks included in our benchmarks balance three priorities. First, they are supported by the MIMIC-III database, the largest publicly repository of rich clinical data currently available. MIMIC-III is already widely used in research and is supported by a large community of researchers, making it ideally suited as the foundation of a benchmark.

Second, our tasks are motivated by important and recognized clinical problems. Bates et al. [6] outline six use cases where big data has the greatest potential to impact the US health care system, including early triage and risk assessment [62], prediction of physiologic decompensation [59], identification of high-cost patients [18], and characterization of complex, multi-system diseases [55]. Each of our prediction tasks is associated with one of these opportunities, albeit with subtle differences due to our use of critical care data.

Finally, our tasks comprise a diversity of classic machine learning problems, including binary and multilabel classification, regression, and time series classification. This last priority led us to omit common clinical prediction tasks that heavily overlapped our final benchmark tasks, such as 30-day and one-year mortality prediction. However, we expect to add these and other tasks to the suite of benchmarks eventually.

2.1 In-hospital mortality

Our first benchmark task involves prediction of in-hospital mortality from data gathered early in an ICU admission. Mortality is a primary outcome of interest in acute care settings, and early detection of at-risk patients is often the key to improving the outcomes. We formulate mortality prediction as a binary classification problem, where the target label indicates whether the patient died before hospital discharge. While traditional models, such as the Simplified Acute Physiology Score (SAPS) [35] include only the first 24 hours, we use a wider 48-hour window to enable the detection of changes and trends that may indicate changes in patient acuity. The PhysioNet/CinC Challenge 2012 used a 48-hour window for the same reason [56]. In contrast to our other tasks, the fixed time window in this problem allows the use of models that accept fixed-length inputs without extensive feature engineering.

Our primary metric for mortality prediction is an area under the receiver operator characteristic curve (AUROC), the most commonly reported metric in mortality prediction research. However, because absolute AUROC scores can be difficult to interpret in the presence of imbalanced classes, we propose two additional metrics that are better suited to problems with imbalanced classes: the area under the precision-recall curve (AUPRC) and minimum of precision and sensitivity (min(Se, $^{+}$P)), a custom metric proposed in the Physionet/CinC Challenge 2012 [56]. In our benchmark data set, 4,493 (10.63%) out of 42,276 patients died in-hospital.

2.2 Decompensation

Our second benchmark involves detection of patients who are physiologically decompensating, or whose conditions are deteriorating precipitously. In an effort to improve outcomes for such patients, many hospitals are deploying "track-and-trigger" programs. For each patient, clinical staff periodically compute an early warning score that quantifies physiologic condition. Low scores, either overall or in certain categories, trigger an alert, which in turn summons a rapid response team of acute care specialists, who then take over care for the triggering patient.

There are a variety of ways to define decompensation, but most objective evaluations of early warning scores are based on accurate prediction of mortality within a fixed time window, e.g., 24 hours [59] after assessment. Following suit, we formulate our decompensation benchmark task as a time series binary classification problem, in which the target label indicates whether the patient will die within the next 24 hours. In our benchmark, predictions are made once per hour for every hour after admission. This problem lends itself to sliding window classifiers and supervised state space models.

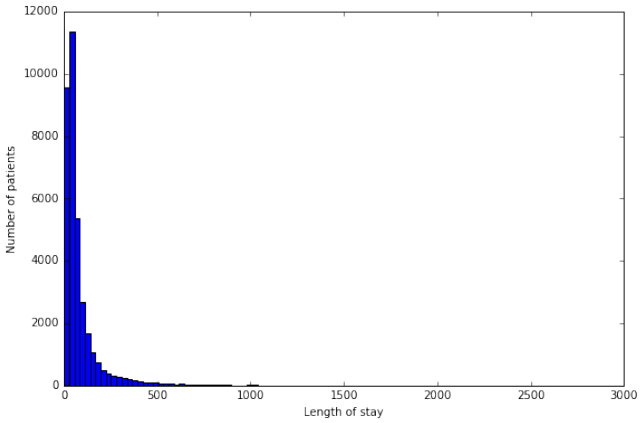
We use the same metrics for decompensation as for mortality, i.e., AUROC, AUPRC, and min(Sen,$^{+}$P). Because we care about per-prediction (vs. per-patient) accuracy in this task, overall performance is computed as the micro-average over all predictions regardless of patient.

2.3 Forecasting length of stay

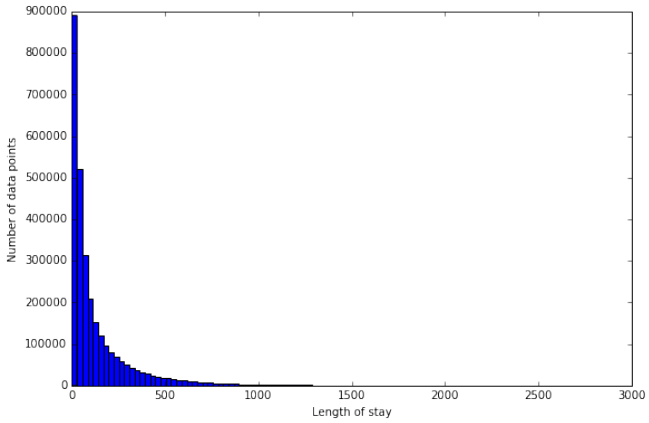
Our third benchmark task involves forecasting hospital length of stay (LOS), one of the most important drivers of overall hospital cost [3, 18]. Patients with long LOS not only require greater hospital resources but also are often among the sickest, with complex, persistent conditions that are not immediately life threatening but nonetheless difficult to treat. One of the major initiatives for reducing health care spending involves early identification of these high-cost patients.

LOS is naturally formulated as a regression task. Traditional research focuses on accurate prediction of LOS early in admission, but in our benchmark we predict the *remaining* length of stay once per hour for every hour after admission, similar to decompensation. Such a model can be used to help hospitals and care units make decisions about staffing and resource on a regular basis, e.g., at the beginning of each day or at shift changes.

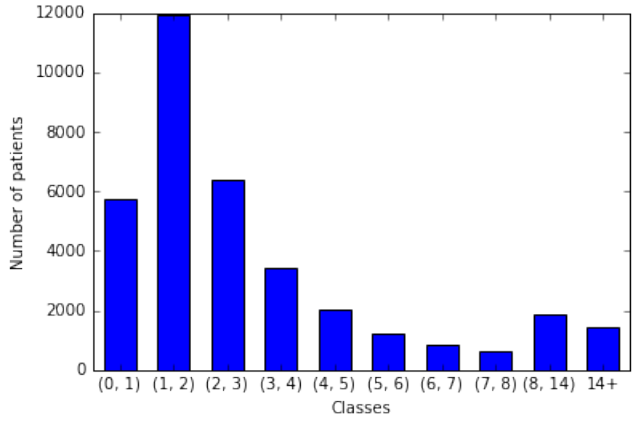
Because there is no widely accepted evaluation metric for LOS predictions, we use standard regression metrics like mean squared



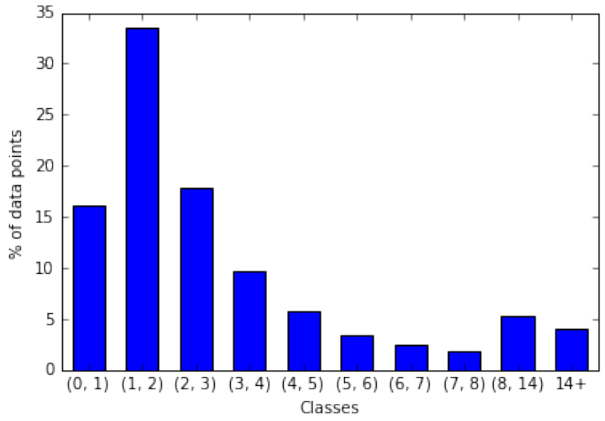
(a) Raw patient length of stay.



(b) Raw daily remaining length of stay.



(c) Bucketed patient length of stay.



(d) Bucketed remaining length of stay.

Figure 1: Distribution of length of stay. Top row: raw (a) patient and (b) daily remaining length of stay. Bottom row: bucketed (c) patient and (d) daily remaining length of stay (less than one day, one each for 1-7 days, between 7 and 14 days, and over 14 days.

error (MSE) and mean absolute percentage error (MAPE). The distribution of patient and daily remaining length of stay are shown in the top row of Figure 1.

In practice, hospitals round to the nearest day when billing, and stays over 1-2 weeks are considered extreme outliers which, if predicted, would trigger special interventions [18]. Thus, we also design a custom metric that captures how length of stay is measured and studied in practice. First, we divide the range of values into ten buckets, one bucket for extremely short visits (less than one day), seven day-long buckets for each day of the first week, and two "outlier" buckets—one for stays of over one week but less than two, and one for stays of over two weeks. This converts length-of-stay prediction into an ordinal multiclass classification problem. To evaluate prediction accuracy for this problem formulation, we use Cohen's linear weighted kappa [7, 15], which measures correlation between ordered items. The bottom row of Figure 1 shows the distribution of patient length of stay across the ten buckets.

2.4 Acute care phenotype classification

Our final benchmark task is the classification of acute care phenotypes: given all data for an ICU stay from admission to discharge, we retrospectively predict which common conditions the patient had.[3] Phenotyping has applications in cohort construction for clinical studies, morbidity detection and risk adjustment, quality improvement and surveillance, and diagnosis [46].

In this task we classify 25 conditions that are common in adult ICUs, including 12 critical (and sometimes life-threatening) conditions, such as respiratory failure and sepsis; 8 chronic conditions that are common comorbidities and risk factors in critical care, such as diabetes and metabolic disorders; and 5 conditions considered "mixed" because they are recurring or chronic with periodic acute

[3]Our decision to perform "retrospective" phenotype classification, rather than early diagnosis prediction, is due in part to a limitation of MIMIC-III: the source of our disease labels, ICD-9 codes, do not have timestamps, so we do not know with certainty when the patient was diagnosed or first became symptomatic. Rather than attempt to assign timestamps using a heuristic, we decided instead to embrace this limitation.

episodes. The full list is shown in Table 1, along with prevalence within the benchmark data set.

Table 1: Common ICU phenotypes along with prevalence within the benchmark data set.

Phenotype	Type	Prev.
Acute and unspecified renal failure	acute	0.21
Acute cerebrovascular disease	acute	0.08
Acute myocardial infarction	acute	0.11
Cardiac dysrhythmias	mixed	0.37
Chronic kidney disease	chronic	0.15
Chronic obstructive pulmonary disease	chronic	0.13
Complications of surgical/medical care	acute	0.27
Conduction disorders	mixed	0.08
Congestive heart failure; nonhypertensive	mixed	0.40
Coronary atherosclerosis and related	chronic	0.53
Diabetes mellitus with complications	mixed	0.14
Diabetes mellitus without complication	chronic	0.19
Disorders of lipid metabolism	chronic	0.29
Essential hypertension	chronic	0.42
Fluid and electrolyte disorders	acute	0.35
Gastrointestinal hemorrhage	acute	0.08
Hypertension with complications	chronic	0.13
Other liver diseases	mixed	0.13
Other lower respiratory disease	acute	0.05
Other upper respiratory disease	acute	0.05
Pleurisy; pneumothorax; pulmonary collapse	acute	0.10
Pneumonia	acute	0.15
Respiratory failure; insufficiency; arrest	acute	0.19
Septicemia (except in labor)	acute	0.24
Shock	acute	0.08

To identify these conditions, we use the single-level definitions from the Health Cost and Utilization (HCUP) Clinical Classification Software (CCS) [2]. These definitions group ICD-9 billing and diagnostic codes into mutually exclusive, largely homogeneous disease categories, reducing some of the noise, redundancy, and ambiguity in the original ICD-9 codes. HCUP CCS code groups are used for reporting to state and national agencies, so they constitute sensible phenotype labels.

Because diseases can co-occur (in fact, 99% of patients in our benchmark data set have more than one diagnosis), we formulate phenotyping as a multi-label classification problem. Similar to Lipton et al. [38], we report macro- and micro-averaged AUROC. We also add a weighted average AUROC metric that takes disease prevalence into account.

3 RELATED WORK

Here we provide a broad review of related work in clinical prediction, clinical applications of deep learning, and multitask learning. There is an extensive body of research on each of these topics, and we will attempt to highlight only the most representative or relevant work since a full treatment is not possible.

3.1 Clinical prediction

Interest in modeling risk of mortality in hospitalized patients dates back over half a century: the Apgar score [5] for assessing risk in newborns was first published in 1952. One of the most widely used risk models is SAPS, which predicts in-hospital mortality from data available within the first 24 hours of admission. Traditionally computed by hand, such scores are designed to require as few inputs as possible and to focus on admission data and individual abnormal observations rather than patterns or trends over time.

Recent research has used machine learning techniques like state space models and time series mining to integrate more detailed data about the patient into mortality prediction. Much of this work aims to make predictions based on complex temporal patterns of physiology rather than individual deranged measurements [40, 42]. Others leverage information from clinical notes, extracted using topic models [9, 23]. All of these approaches outperform baselines based on predefined scores, but results are generally not comparable due to use of different data.

The detection of patients that are physiologically decompensating, or whose conditions are deteriorating rapidly, is at the center of "track-and-trigger" initiatives, which aim to improve outcomes by rapidly delivering specialized care to the sickest patients. Early warning scores, such as the National Early Warning Score (NEWS) [59] being deployed throughout the United Kingdom, detect decompensation using clinician-defined thresholds on a handful of physiologic variables. Like traditional risk scores, most early warning scores are applied manually and emphasize simplicity.

There has been a great deal of machine learning research on the related problem of condition monitoring, but most of this work formulates the task as anomaly detection [4, 41, 52], rather than continuous mortality prediction. Other related works model the risk of individual conditions such as *Clostridium difficile* infection [58] or sepsis [27] over time. Caballero Barajas and Akella [9] and Ghassemi et al. [22] both train temporal risk models but use them to forecast eventual, rather than imminent, mortality.

Hospitals use patient LOS as both a measure of a patient's acuity and for scheduling and resource management [3]. Most LOS research has focused on identifying factors that influence LOS [28] rather than building predictive models. Both severity of illness scores [47] and early warning scores [48] have been used to predict LOS but with mixed success. There has been limited machine learning research concerned with length of stay, most of it focused on specific conditions [50] and cohorts [24]. None of this work has addressed continuous prediction of length of stay over time.

Phenotype classification is a relatively new medical informatics problem that has seen a surge of interest from machine learning researchers [1, 25, 38, 42]. Recent work has explored ways to make development of phenotype classifiers more scalable by reducing the need for large numbers of labeled training records [1, 25]. Lipton et al. [38] were among the first researchers to formulate phenotyping as a multi-label time series classification problem, which is the setup we use in this work.

3.2 Deep learning for clinical data

Because our proposed framework leverages techniques from deep learning, here we call special attention to both historical and recent

research that applies neural network architectures to clinical data. In predicting mortality from early admission data, feedforward neural networks nearly always outperform baselines based on logistic regression or severity of illness scores [10, 11, 14]. However, to our knowledge, there has been little work on using recurrent neural networks to analyze temporal clinical data and model risk of mortality in hospitalized patients. Likewise, neural networks have not been widely applied to detection of patient decompensation or continuous prediction of imminent mortality.

There is a great deal of early research that uses neural networks to predict LOS in hospitalized patients [24, 43]. However, rather than regression, much of this work formulates the task as binary classification aimed at identifying patients at risk for long stays [8]. Recently, novel deep learning architectures have been proposed for survival analysis [53, 61], a similar time-to-event regression task with right censoring.

Phenotyping has been a popular application for deep learning researchers in recent years, though model architecture and problem definition vary widely. Lasko et al. [34] and Che et al. [12] each applied feedfoward architectures to clinical time series in sliding window fashion in order to mine phenotypic patterns of physiology. Razavian et al. [54] proposed a novel temporal convolutional network for a related problem, predicting future diagnoses. Choi et al. [13] used a long short-term memory network to model sequences of diagnostic codes, a proxy task for disease progression.

Our work is closely related to several recent papers. Lipton et al. [38] were the first to show that recurrent neural networks could classify over a hundred different diagnoses in variable length clinical time series. Purushotham et al. [51] use an adversarial recurrent framework to perform domain adaptation between differnt patient age groups They apply this model to both mortality prediction and diagnostic classification, but use separate data sets for each task.

3.3 Multitask learning

Multitask learning has its roots in clinical prediction: Caruana et al. [10] used future lab values as auxiliary targets during training improved prediction of mortality among pneumonia patients. Much of the subsequent work on multitask learning found applications in health care and related domains, where correlations between related tasks can improve prediction performance – especially with limited training data. Nori et al. [45] and Che et al. [12] both encode predefined similarity between diseases (based on an ontology) into an explicit graph Laplacian regularizer for mortality prediction and phenotype classification. Lipton et al. [38] and Razavian et al. [54] formulate phenotyping as a multi-label classification problem and use neural networks to implicitly capture co-morbidities in hidden layers.

After Caruana et al. [10], there has been only limited work on multitask learning with heterogeneous tasks. Yang et al. [60] introduce a general framework for learning a mixture of regression and classification models with joint feature selection and apply it to gene expression data. Ngufor et al. [44] apply a similar framework to jointly solving thirteen related clinical tasks, including predicting mortality and length of stay. In computer vision, Li et al. [37] and Elhoseiny et al. [19] both showed that using a single model to perform joint detection (classification) and pose estimation (regression) improves generalization on both tasks. However, none of this work addresses problem settings where sequential or temporal structure varies across tasks. The closest work in spirit to ours is by Collobert and Weston [16], who use a single convolutional network to perform a variety of natural language tasks (part-of-speech tagging, named entity recognition, and language modeling) with diverse sequential structure.

4 METHODS

We begin by introducing notation and briefly revisiting the fundamentals of long short-term memory recurrent neural networks [29], or LSTM RNNs, in order to establish our basic model architecture. We then describe how to build a unified heterogeneous multitask LSTM for joint prediction of all our clinical tasks. For a given ICU stay of length T hours, we assume we are given a series of regularly spaced observations $\{x_t\}_{t\geq 1}^{T}$, where x_t is a vector of clinical observations at hour t.

We also have a set of targets for each stay: $\{d_t\}_{t\geq 1}^{T}$ where $d_t \in \{0, 1\}$ is a list of T binary labels for decompensation, one for each hour; $m \in \{0, 1\}$ is single binary label indicating whether the patient died in-hospital; $\{\ell_t\}_{t\geq 1}^{T}$ where $\ell_t \in \mathbb{R}$ is a list of real valued numbers indicating *remaining* length of stay (hours until discharge) at each time step; and $p_{1:K} \in \{0, 1\}^K$ is a vector of K binary phenotype labels. When training our models to predict length of stay, we instead use a set of categorical labels $\{l_t\}_{t\geq 1}^{T}$ where $l_t \in \{1, \ldots, 10\}$ indicates in which of the ten length-of-stay buckets ℓ_t belongs. When used in the context of equations (e.g., as the output of a softmax or in a loss function), we will interpret l_t as a one-of-ten hot binary vector, indexing the ith entry as l_{ti}.

The LSTM takes a sequence $\{x_t\}_{t\geq 1}^{T}$ of length T as its input and outputs a T-long sequence of $\{h_t\}_{t\geq 1}^{T}$ hidden state vectors using the following equations:

$$
\begin{aligned}
i_t &= \sigma(x_t W_{xi} + h_{t-1} W_{hi} + w_{ci} \odot c_{t-1} + b_i) \\
f_t &= \sigma(x_t W_{xf} + h_{t-1} W_{hf} + w_{cf} \odot c_{t-1} + b_f) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(x_t W_{xc} + h_{t-1} W_{hc} + b_c) \\
o_t &= \sigma(x_t W_{xo} + h_{t-1} W_{ho} + w_{co} \odot c_t + b_o) \\
h_t &= o_t \odot \sigma_h(c_t)
\end{aligned}
$$

The σ (sigmoid) and tanh functions are applied element-wise. Unlike Lipton et al. [38], we use peephole connections (w_{ci}, w_{cf} and w_{co}) [21]. The W matrices and w and b vectors are the trainable parameters of the LSTM. Later we will use $h_t = LSTM(x_t, h_{t-1})$ as a shorthand for the above equations.

We then add the following output layers to predict each of our benchmark tasks: $\left\{\widehat{d}_t\right\}_{t\geq 1}^{T}$, $\widehat{m}$, $\left\{\widehat{l}_t\right\}_{t\geq 1}^{T}$, $\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_K$:

$$
\begin{aligned}
\widehat{d}_t &= \sigma\left(w^{(d)} h_t + b^{(d)}\right) \\
\widehat{m} &= \sigma\left(w^{(m)} h_{48} + b^{(m)}\right) \\
\widehat{l}_t &= \text{softmax}\left(w^{(l)} h_t + b^{(l)}\right) \\
\widehat{p}_i &= \sigma\left(W_{i,\cdot}^{(p)} h_T + b_i^{(p)}\right)
\end{aligned}
$$

For in-hospital mortality, we consider only the first 48 timesteps $\{x_t\}_{t\geq 1}^{t_m}$, and predict in-hospital mortality $\widehat{m}$ at $t_m = 48$ by adding a multilayer perceptron output layer with sigmoid activation that takes only h_{t_m} as its input. For decompensation, we consider the full sequence $\{x_t\}_{t\geq 1}^{T}$ and generate a sequence of mortality predictions $\{\widehat{d}\}_{t\geq 1}^{T}$ by adding an MLP output at every step. For phenotyping, we consider the full sequence but predict phenotypes $\widehat{p}$ only at the last timestep T by adding a multi-label MLP with sigmoid activations only at the last timestep.

Similar to decompensation, we predict LOS by adding an MLP output at each timestep, but we experimented with two different types of output layers. The first is a dense linear output MLP that makes a real valued prediction $\widehat{\ell}_t$. The second is a softmax output MLP with ten neurons that outputs a distribution over the ten LOS buckets $\widehat{l}_t$.

For all binary classification problems (mortality, decompensation, and phenotyping), we use cross entropy loss defined as

$$\text{CE}(y, \widehat{y}) = -(y \cdot \log(\widehat{y}) + (1-y) \cdot \log(1-\widehat{y}))$$

For predicting LOS buckets, we use multiclass cross entropy defined over the B buckets:

$$\text{MCE}(y, \widehat{y}) = -\sum_{k=1}^{B} y_k \log(\widehat{y}_k)$$

For raw LOS, we use MSE. For decompensation and LOS, we average the loss over steps. We summarize the per-task loss functions for a single example as follows:

$$loss_d = \frac{1}{T}\sum_{t=1}^{T} \text{CE}(d_t, \widehat{d}_t)$$

$$loss_\ell = \frac{1}{T}\sum_{t=1}^{T} \left(\ell_t - \widehat{\ell}_t\right)^2$$

$$loss_l = \frac{1}{T}\sum_{t=1}^{T} \text{MCE}(l_{tk}, \widehat{l}_{tk})$$

$$loss_m = \text{CE}(m, \widehat{m})$$

$$loss_p = \frac{1}{K}\sum_{i=1}^{K} \text{CE}(p_k, \widehat{p}_k)$$

The full multitask LSTM architecture is illustrated in Fig. 2. In our experiments we found that the categorical formulation of length of stay ($\widehat{l}_t$) performed better than the raw version ($\widehat{\ell}_t$) and so we used $\widehat{l}_t$ in the final multitask architecture. To train our multitask LSTM, we used the following weighted loss function:

$$loss_{mt} = \lambda_d \cdot loss_d + \lambda_l \cdot loss_l + \lambda_m \cdot loss_m + \lambda_p \cdot loss_p$$

Following [49] we apply dropout on non-recurrent connections between LSTM layers and before outputs.

5 DATA

We built our benchmarks using data from the MIMIC-III critical care database, which includes over 60,000 ICU stays across 40,000 critical care patients. We excluded all neonatal and pediatric patients (age 18 or younger at time of ICU stay) because the physiology of pediatric critical care patients differs significantly from adults. We also excluded any hospital admission with multiple ICU stays or transfers between different ICU units or wards. The final cohort has 33,798 unique patients with a total of 42,276 hospital admissions and ICU stays. To facilitate direct comparison of model performance across different projects, we defined a fixed test set of 5,070 (15%) patient stays. However, we encourage researchers using our benchmarks to follow best practices by avoiding looking at the test data as much as possible.

We determined in-hospital mortality by comparing patient date of death (DOD) with hospital admission and discharge times. The mortality rate within the benchmark cohort is 10.63%. We computed decompensation labels in sliding window fashion, based on whether patient DOD falls within the next 24 hours from the current time. For LOS, we used the existing LOS field from the MIMIC-III database and then computed the hourly remaining LOS targets using a sliding window. Figure 1 shows the distributions of patient LOS and hourly remaining LOS in our final cohort.

We determined phenotype labels based on the MIMIC-III ICD-9 diagnosis table. First, we mapped each code to its HCUP CCS category, retaining only the 25 categories from Table 1. We then matched diagnoses to ICU stays using the hospital admission identifier, since ICD-9 codes in MIMIC-III are associated with hospital visits, not ICU stays. By excluding hospital admissions with multiple ICU stays, we reduced some of the ambiguity in these labels: there is only one ICU stay per hospital admission with which the diagnosis can be associated. While a certain amount of label noise is inevitable in ICD-9 codes, by taking these steps (as well as applying CCS groupings), we are reasonably confident that we have significantly reduced the amount of noise in our final phenotype labels.

As of this writing, our benchmark data set includes 17 physiologic time series variables that represent a subset of the variables included in the Physionet/CinC Challenge 2012 data set [56]. Code for generating our exact benchmark data set from MIMIC-III is available at https://github.com/yerevann/mimic3-benchmarks.

6 EXPERIMENTS

During model development, we use 15% of the predefined training set as validation data and train the models on the remaining 85%. All architectures are trained for several configurations of hyperparameters, and the best model is chosen according to the performance on the validation set. The final scores are reported on the test set, which we used sparingly during model development in order to avoid unintentional test set leakage.

6.1 Baselines

We compare our multitask LSTM with two types of single-task baselines: linear (logistic) regression with hand-engineered features and single-task LSTMs, both of which have previously been effective in modeling clinical time series [38, 51]. We omit baselines based on severity of illness scores like SAPS for two reasons. First, in a variety of previous work, strong machine learning models were shown to outperform SAPS by large margins [9, 23, 56]. Second, the current version of our benchmark data set does not include all of the SAPS variables, so any comparison would not be apples-to-apples.

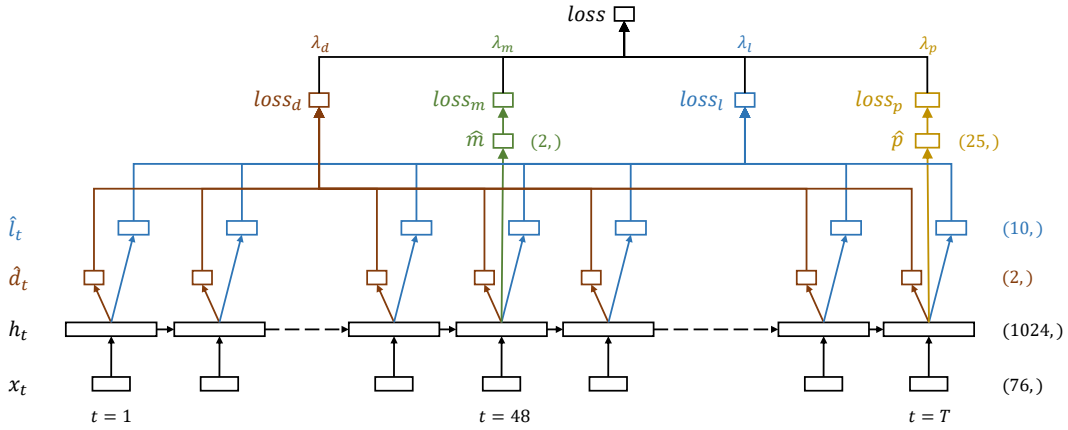


Figure 2: LSTM-based network architecture for multitask learning.

For our linear baselines, we use a more elaborate version of the hand-engineered features described in Lipton et al. [38]: for each variable, we compute six different sample statistic features on seven different subsequences of a given time series. For mortality, this includes only the first 48 hours, while for phenotyping this includes the entire ICU stay. For decompensation and LOS, we apply the classifier in sliding window fashion, so that the features are recomputed at each time step. The per-subsequence features include minimum, maximum, mean, standard deviation, skew and number of measurements. The seven subsequences include the full time series, the first 10% of time, first 25% of time, first 50% of time, last 50% of time, last 25% of time, last 10% of time. In total, we obtain $17 \times 7 \times 6 = 714$ features per time series.

We train a separate logistic regression classifier for each of mortality, decompensation, and the 25 phenotypes. For LOS, we trained a softmax regression model to solve the 10-class bucketed LOS problem.

For LSTMs we resample the time series into regularly spaced intervals. If there are multiple measurements of the same variable in the same interval, we use the value of the last measurement. We impute the missing values using the previous value if it exists and a pre-specified "normal" value otherwise.[4] In addition, we also as a binary mask input for each variable indicating the timesteps that contain a true (vs. imputed) measurement [39]. Categorical variables are encoded using a one-hot vector at each timestep. Then the inputs are then normalized by subtracting the mean and dividing by standard deviation. Statistics are calculated per variable after imputation of missing values.

6.2 Model selection

We used a grid search to tune all hyperparameters based on validation set performance. For logistic regression models, we tried both L1 and L2 regularization with different coefficients. For mortality and decompensation, the best performing logistic regression used L2 regularization with $C = 0.001$. For phenotyping, the best performing logistic regression used L1 regularization with $C = 0.1$. For LOS, the best performing logistic regression used L2 regularization with $C = 10^{-5}$.

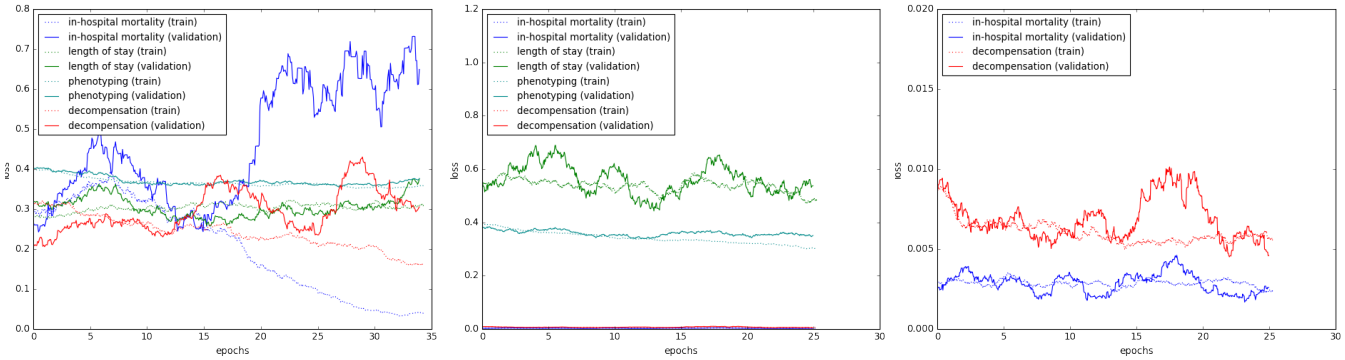
For single-task LSTMs, hyperparameters included dropout, numbers of memory cells in LSTM layers, and whether to use one or two LSTM layers. All LSTMs were trained using ADAM [**?**] with a 10^{-4} learning rate and $\beta_1 = 0.5$. The best single-task LSTM architecture for mortality and decompensation had a single LSTM layer with 256 neurons and no dropout. Phenotyping was the only task where adding a second LSTM layer and dropout helped to improve performance. The best LOS LSTM model had one layer of 256 neurons and no dropout. The best performing 1-layer and 2-layer phenotype LSTM models used 1024 and 512 neurons per layer, respectively. Table 5 reports metrics for both versions of LSTM.

For the multitask LSTM, we used only a single LSTM layer with 1024 neurons and used grid search to tune the values of λ_d, λ_m, λ_l and λ_p in the weight loss function. Overfitting was a serious problem, with mortality and decompensation prediction validation performance degrading faster than the others. To ameliorate this, we used lower values for λ_d and λ_m. We obtained our best results using the following settings: $\lambda_d = 0.1, \lambda_m = 0.02, \lambda_l = 0.5, \lambda_p = 1.0$. The scores for this model are reported in the tables under the name "Multitask LSTM."

6.3 Results

The results for each of the mortality, decompensation, LOS, and phenotyping tasks are reported in Table 2, Table 3, Table 4, and Table 5, respectively. We first note that LSTM-based models outperformed linear models by substantial margins across across all metrics on every task. This is consistent with previous research comparing neural networks to linear models for mortality prediction [14] and phenotyping [38], but it is nonetheless noteworthy because questions still remain about the potential effectiveness of deep learning for health data, especially given the often modest size of the data relative to their complexity. Our results provide further evidence that complex architectures can be effectively trained on non-Internet scale health data and that while challenges like overfitting persist, they can be mitigated with careful regularization schemes, including dropout and multitask learning.

[4]The normal values we used can be found at https://github.com/YerevaNN/mimic3-benchmarks/blob/master/resources/variable_ranges.csv.



(a) Multitask LSTM with equal task weights.

(b) Best overall multitask LSTM.

(c) Best overall multitask LSTM.

Figure 3: Convergence of per-task training and validation losses in multitask LSTMs. (a) Multitask LSTM with equal task weights. (b) LOS and phenotyping and (c) mortality and decompensation losses for best overall multitask LSTM.

The multitask LSTM clearly outperformed the baselines on the mortality and decompensation tasks. For LOS we see mixed results: the single-task LSTM has a slight advantage in terms of Cohen's linear kappa score (for the bucketed evaluation), but the differences are unlikely to be significant. On the other hand, the multitask LSTM outperforms single-task LSTM on the standard regression metrics.

Likewise, on the phenotyping task, the final multitask LSTM falls short of the best one- and two-layer single-task LSTMs. We hypothesize that this is because phenotype classification is itself a multitask problem and already benefits from regularization by sharing LSTM layers across the 25 different phenotypes. The addition of further tasks with loss weighting may limit the multitask LSTM's ability to effectively learn to recognize individual phenotypes. Table 6 shows the per-phenotype AUROC for the two-layer single-task LSTM. We observe that performance on the individual diseases varies widely from 0.6779 (essential hypertension) and 0.9089 (acute cerebrovascular disease).

While training the multitask LSTM, we observed that validation performance for different tasks converged and then overfit at different rates. This is illustrated in Figure 3a. Because we chose the best multitask LSTM based on overall validation performance across all tasks, it is unlikely that the chosen model achieved optimal performance on each individual task. This puts the multitask LSTM at a significant disadvantage when compared against single-task LSTMs, each which optimizes individual task performance.

This also meant that individual task performance was highly sensitive to choice of loss weights in our weighted multitask loss function. In particular, we found that by varying the loss weights, we could trade off mortality and decompensation accuracy against LOS and phenotyping performance. As shown in Figure 3a, the multitask LSTM with equally weighted losses quickly overfits the mortality and decompensation problems but makes only limited progress on LOS and phenotyping. On the other hand, an LSTM with loss weights $\lambda_d = 1.0$, $\lambda_m = 0.1$, $\lambda_l = 0.1$, and $\lambda_p = 1.0$ suffers less overfitting and makes gradual progress on all tasks. This

Table 2: Results for in-hospital mortality prediction task

Model	AUROC	AUPRC	min(Se, $^+$P)
Logistic regression	0.8442	0.4717	0.4693
LSTM	0.8540	0.5164	0.4905
Multitask LSTM	0.8550	0.4926	0.4745
Multitask LSTM*	0.8625	0.5169	0.4987

Table 3: Results for decompensation prediction task

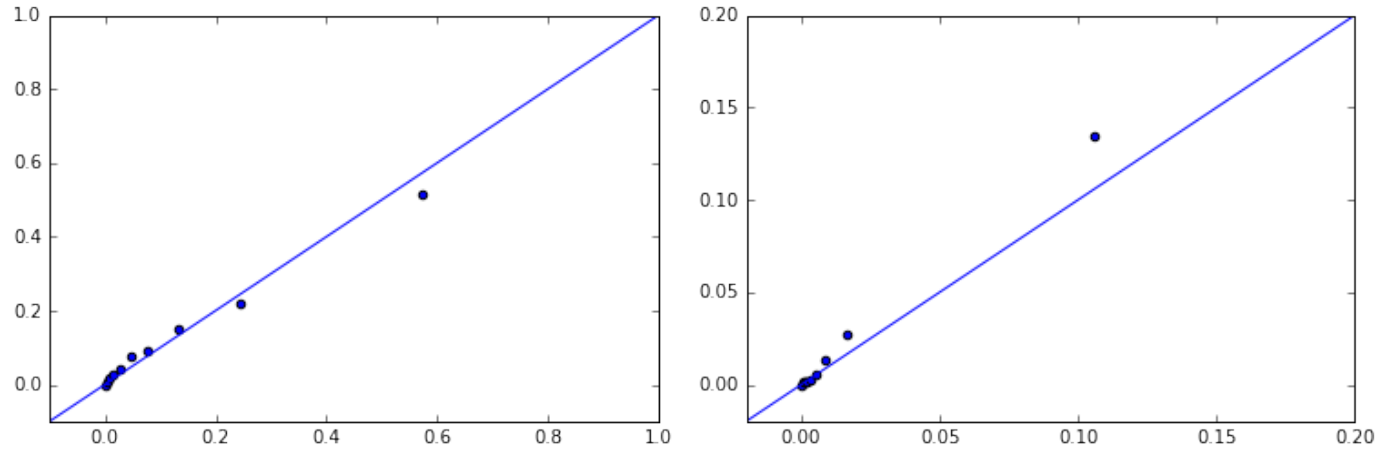
Model	AUROC	AUPRC	min(Se, $^+$P)
Logistic regression	0.8704	0.2132	0.2688
LSTM	0.8946	0.2980	0.3438
Multitask LSTM	0.9004	0.3192	0.3484
Multitask LSTM**	0.9119	0.3322	0.3593

Table 4: Results for length of stay prediction task

Model	Kappa	MSE	MAPE
Logistic regression	0.4021	63385	573.5
LSTM	0.4266	42165	235.9
Multitask LSTM	0.4258	42131	188.5

model achieved 0.8625 AUROC for in-hospital mortality prediction (reported as "Multitask LSTM*" in Table 2). Another model with $\lambda_d = 1.0$, $\lambda_m = 0.1$, $\lambda_l = 0.1$ and $\lambda_p = 1.0$ got an 0.9119 AUC score for decompensation prediction task (reported as "Multitask LSTM**" in Table 3).

For risk-related tasks like mortality and decompensation, we are also interested how reliable the probabilities estimated by our predictive models are. This is known as *calibration* and is a common method for evaluating predictive models in the clinical research



(a) In-hospital mortality. **(b) Decompensation.**

Figure 4: Calibration of in-hospital (a) mortality and (b) decompensation predictions by best overall multitask LSTM. Plots predicted vs. actual probability of mortality.

Table 5: Results for phenotyping task

Model	micro AUC	macro AUC	weighted AUC
Logistic regression	0.8007	0.7408	0.7320
1-layer LSTM	0.8206	0.7701	0.7573
2-layer LSTM	0.8213	0.7707	0.7587
Multitask LSTM	0.8174	0.7661	0.7533

literature. In a well calibrated model, 10% all patients who receive a predicted probability of decompensation do in fact decompensate.

We included no formal measure of calibration in our benchmark evaluations, but in Figure 4 we informally visualize calibration for mortality and decompensation predictions from our best overall multitask LSTM using reliability plots. These are scatter plots of predicted probability (computed by creating decile bins of predictions and then taking the mean value within each bin) vs. actual probability (the rate of, e.g., mortality, within each bin). Better calibrated predictions will fall closer to the diagonal. We see that the LSTM mortality predictions look reasonably calibrated, but the LSTM appears to unpredict probability of decompensation.

7 CONCLUSION

In this paper, we proposed four first-of-their-kind standardized benchmarks for machine learning researchers interested in clinical data problems, including in-hospital mortality, decompensation, length of stay, and phenotyping. Our benchmark data set is similar to other MIMIC-III patient cohorts described in by machine learning publications but makes use of a larger number of patients and is immediately accessible to other researchers who wish to replicate our experiments or build upon our work.

We also described a unified neural net architecture for learning and predicting all four clinical tasks simultaneously. This model consists of a heterogeneous multitask LSTM RNN and a custom loss function that consists of a weighted combination of the individual task losses. To our knowledge, this is the first multitask RNN architecture designed for tasks that have both heterogeneous types

Table 6: Per-phenotype classification performance for best overall multitask LSTM.

Phenotype	AUROC
Acute and unspecified renal failure	0.8035
Acute cerebrovascular disease	0.9089
Acute myocardial infarction	0.7695
Cardiac dysrhythmias	0.684
Chronic kidney disease	0.7771
Chronic obstructive pulmonary disease	0.6786
Complications of surgical/medical care	0.7176
Conduction disorders	0.726
Congestive heart failure; nonhypertensive	0.7608
Coronary atherosclerosis and related	0.7922
Diabetes mellitus with complications	0.8738
Diabetes mellitus without complication	0.7897
Disorders of lipid metabolism	0.7213
Essential hypertension	0.6779
Fluid and electrolyte disorders	0.7405
Gastrointestinal hemorrhage	0.7413
Hypertension with complications	0.76
Other liver diseases	0.7659
Other lower respiratory disease	0.688
Other upper respiratory disease	0.7599
Pleurisy; pneumothorax; pulmonary collapse	0.7027
Pneumonia	0.8082
Respiratory failure; insufficiency; arrest	0.9015
Septicemia (except in labor)	0.8426
Shock	0.876

and varied temporal structure. We found that the weighted loss function improves overall performance by ensuring that the model makes reasonable progress on all learning tasks.

Our results demonstrate that the phenotyping and length-of-stay prediction tasks are more challenging and require larger model architectures than mortality and decompensation prediction tasks. Even small LSTM models easily overfit the latter two problems. We also demonstrated that the proposed multitask learning architecture allows us to extract certain useful information from the input sequence that single-task models could not leverage, which explains the better performance of multitask LSTM on the easier tasks. We did not, however, find any significant benefit in using multitask learning for the more difficult tasks.

We plan to continue improving and expanding our benchmark data set by adding further observational variables, inputs and outputs, and medications and treatments. This should not only enable improved performance on the existing benchmarks but also facilitate interesting new research problems, such as treatment recommendation and causal inference. We will also add results for additional baselines, both traditional (SAPS) and modern (Gaussian Processes) to future revisions of this manuscript and post and track published results using our benchmark at project github repository, [5] where code to construct the benchmark and train and run our models can be found. We also plan to refine our evaluations based on community feedback. For example, we would like to include formal measures calibration (visualized informally in Figure 4) for mortality and decompensation based on, e.g., the Hosmer-Lemeshow test.

We are interested in further investigating the practical challenges of multitask training. In particular, for our four very different tasks, the model converges and then overfits at very different rates during training. This is often addressed through the use of heuristics, including a multitask variant of early stopping, in which we identify the best epoch for each task based on individual task validation loss.[6] We proposed the use of per-task loss weighting, which reduced the problem but did not fully mitigate it. One promising direction dynamically adapt these coefficients during training, similar to the adaptation of learning rates in optimizers like ADAM [33].

REFERENCES

[1] Vibhu Agarwal, Tanya Podchiyska, Juan M Banda, Veena Goel, Tiffany I Leung, Evan P Minty, Timothy E Sweeney, Elsie Gyang, and Nigam H Shah. 2016. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association* 23, 6 (2016), 1166.

[2] Agency for Healthcare Research and Quality. 2012. Clinical Classifications Software (CCS) for ICD-9-CM Fact Sheet. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp. (2012). Accessed February 9, 2017.

[3] Agency for Healthcare Research and Quality. 2014. Selecting Quality and Resource Use Measures: A Decision Guide for Community Quality Collaboratives. https://www.ahrq.gov/professionals/quality-patient-safety/quality-resources/tools/perfmeasguide/perfmeaspt3.html. (2014). Accessed February 9, 2017.

[4] Norm Aleks, Stuart J Russell, Michael G Madden, Diane Morabito, Kristan Staudenmayer, Mitchell Cohen, and Geoffrey T Manley. 2009. Probabilistic detection of short events, with application to critical care monitoring. In *Advances in Neural Information Processing Systems*. 49–56.

[5] Virgina Apgar. 1952. A proposal for a new method of evaluation of the newborn. *Curr. Res. Anesth. Analg.* 32, 449 (1952), 260–267.

[6] David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs* 33, 7 (2014), 1123–1131.

[7] Robert L Brennan and Dale J Prediger. 1981. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement* 41, 3 (1981), 687–699.

[8] Timothy G Buchman, Kenneth L Kubos, Alexander J Seidler, and Michael J Siegforth. 1994. A comparison of statistical and connectionist models for the prediction of chronicity in a surgical intensive care unit. *Critical care medicine* 22, 5 (1994), 750–762.

[9] 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.

[10] Rich Caruana, Shumeet Baluja, Tom Mitchell, and others. 1996. Using the future to "sort out" the present: Rankprop and multitask learning for medical risk evaluation. In *Advances in Neural Information Processing Systems (NIPS) 8*. 959–965.

[11] Leo Anthony Celi, Sean Galvin, Guido Davidzon, Joon Lee, Daniel Scott, and Roger Mark. 2012. A database-driven decision support system: customized mortality prediction. *Journal of personalized medicine* 2, 4 (2012), 138–148.

[12] 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.

[13] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *1st Machine Learning for Healthcare Conference*.

[14] Gilles Clermont, Derek C Angus, Stephen M DiRusso, Martin Griffin, and Walter T Linde-Zwirble. 2001. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic fmultion models. *Critical care medicine* 29, 2 (2001), 291–296.

[15] J Cohen. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurrement, 20, 37-46. *I960* (1960).

[16] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, 160–167.

[17] Healthcare Cost, Utilization Project (HCUP), and others. 2014. Introduction to the HCUP National Inpatient Sample (NIS) 2012. *Agency for Healthcare Research and Quality, Rockville, MD* (2014).

[18] Deborah Dahl, Greg G Wojtal, Michael J Breslow, Debra Huguez, David Stone, and Gloria Korpi. 2012. The high cost of low-acuity ICU outliers. *Journal of Healthcare Management* 57, 6 (2012), 421–433.

[19] Mohamed Elhoseiny, Tarek El-Gaaly, Amr Bakry, and Ahmed Elgammal. 2016. A Comparative Analysis and Study of Multiview CNN Models for Joint Object Categorization and Pose Estimation. In *Proceedings of The 33rd International Conference on Machine Learning*. 888–897.

[20] David Ferrucci, Anthony Levas, Sugato Bagchi, David Gondek, and Erik T Mueller. 2013. Watson: beyond jeopardy! *Artificial Intelligence* 199 (2013), 93–105.

[21] Felix A Gers and Jürgen Schmidhuber. 2000. Recurrent nets that time and count. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, Vol. 3. IEEE, 189–194.

[22] 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.

[23] 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 *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, Vol. 2015. NIH Public Access, 446.

[24] Jim Grigsby, Robert Kooken, and John Hershberger. 1994. Simulated neural networks to predict outcomes, costs, and length of stay among orthopedic rehabilitation patients. *Archives of physical medicine and rehabilitation* 75, 10 (1994), 1077–1081.

[25] Yoni Halpern, Steven Horng, Youngduck Choi, and David Sontag. 2016. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association* 23, 4 (2016), 731.

[26] JaWanna Henry, Yuriy Pylypchuk, MPA Talisha Searcy, and Vaishali Patel. 2015. Adoption of Electronic Health Record Systems among US Non-Federal Acute Care Hospitals: 2008-2015. *ONC Data Brief* 35 (2015).

[27] 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.

[5] https://github.com/yerevann/mimic3-benchmarks

[6] Recommended in private communications by Rich Caruana of Microsoft Research.

[28] Thomas L Higgins, William T McGee, Jay S Steingrub, John Rapoport, Stanley Lemeshow, and Daniel Teres. 2003. Early indicators of prolonged intensive care unit stay: Impact of illness severity, physician staffing, and pre–intensive care unit length of stay. *Critical care medicine* 31, 1 (2003), 45–51.

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

[30] ImageNet. 2016. ImageNet Large Scale Visual Recognition Challenge Website. http://imagenet.org. (2016). Accessed Feburary 8, 2017.

[31] Kenneth V Iserson and John C Moskop. 2007. Triage in medicine, part I: concept, history, and types. *Annals of emergency medicine* 49, 3 (2007), 275–281.

[32] 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).

[33] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[34] Thomas A. Lasko, Joshua C. Denny, and Mia A. Levy. 2013. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLoS ONE* 8, 6 (06 2013), e66341.

[35] Jean-Roger Le Gall, Philippe Loirat, Annick Alperovitch, Paul Glaser, Claude Granthil, Daniel Mathieu, Philippe Mercier, Remi Thomas, and Daniel Villers. 1984. A simplified acute physiology score for ICU patients. *Critical care medicine* 12, 11 (1984), 975–977.

[36] Joon Lee and David M Maslove. 2017. Customization of a severity of illness score using local electronic medical record data. *Journal of intensive care medicine* 32, 1 (2017), 38–47.

[37] Sijin Li, Zhi-Qiang Liu, and Antoni B Chan. 2014. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 482–489.

[38] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. 2016. Learning to Diagnose with LSTM Recurrent Neural Networks. *International Conference on Learning Representations* (2016).

[39] Zachary C Lipton, David C Kale, and Randall Wetzel. 2016. Modeling Missing Data in Clinical Time Series with RNNs. *Machine Learning for Healthcare Conference* (2016).

[40] Yuan Luo, Yu Xin, Rohit Joshi, Leo Celi, and Peter Szolovits. 2016. Predicting ICU Mortality Risk by Grouping Temporal Trends from a Multivariate Panel of Physiologic Measurements. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[41] Yi Mao, Wenlin Chen, Yixin Chen, Chenyang Lu, Marin Kollef, and Thomas Bailey. 2012. An integrated data mining approach to real-time clinical monitoring and deterioration warning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1140–1148.

[42] Benjamin M Marlin, David C Kale, Robinder G Khemani, and Randall C Wetzel. 2012. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. ACM, 389–398.

[43] Bert A Mobley, Renee Leasure, and Lynda Davidson. 1995. Artificial neural network predictions of lengths of stay on a post-coronary care unit. *Heart & Lung: The Journal of Acute and Critical Care* 24, 3 (1995), 251–256.

[44] Che Ngufor, Sudhindra Upadhyaya, Dennis Murphree, Daryl Kor, and Jyotishman Pathak. 2015. Multi-task learning with selective cross-task transfer for predicting bleeding and other important patient outcomes. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*. IEEE, 1–8.

[45] Nozomi Nori, Hisashi Kashima, Kazuto Yamashita, Hiroshi Ikai, and Yuichi Imanaka. 2015. Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 855–864.

[46] Anika Oellrich, Nigel Collier, Tudor Groza, Dietrich Rebholz-Schuhmann, Nigam Shah, Olivier Bodenreider, Mary Regina Boland, Ivo Georgiev, Hongfang Liu, Kevin Livingston, and others. 2015. The digital revolution in phenotyping. *Briefings in bioinformatics* (2015), bbv083.

[47] Turner M Osler, Frederick B Rogers, Laurent G Glance, Myra Cohen, Robert Rutledge, and Steven R Shackford. 1998. Predicting survival, length of stay, and cost in the surgical intensive care unit: APACHE II versus ICISS. *Journal of Trauma and Acute Care Surgery* 45, 2 (1998), 234–238.

[48] R Paterson, DC MacLeod, D Thetford, A Beattie, C Graham, S Lam, and D Bell. 2006. Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit. *Clinical Medicine* 6, 3 (2006), 281–284.

[49] Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. 2014. Dropout improves recurrent neural networks for handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, 285–290.

[50] Walter E Pofahl, Steven M Walczak, Ethan Rhone, and Seth D Izenberg. 1998. Use of an artificial neural network to predict length of stay in acute pancreatitis. *The American Surgeon* 64, 9 (1998), 868.

[51] Sanjay Purushotham, Wilka Carvalho, Tanachat Nilanon, and Yan Liu. 2017. Variational Recurrent Adversarial Deep Domain Adaptation. *International Conference on Learning Representations* (2017).

[52] John A Quinn, Christopher KI Williams, and Neil McIntosh. 2009. Factorial switching linear dynamical systems applied to physiological condition monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 9 (2009), 1537–1551.

[53] Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. 2016. Deep Survival Analysis. In *1st Machine Learning for Healthcare Conference*.

[54] Narges Razavian, Jake Marcus, and David Sontag. 2016. Multi-task prediction of disease onsets from longitudinal lab tests. In *1st Machine Learning for Healthcare Conference*.

[55] Suchi Saria and Anna Goldenberg. 2015. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems* 30, 4 (2015), 70–75.

[56] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark. 2012. Predicting in-hospital mortality of ICU patients: The PhysioNet/Computing in cardiology challenge 2012. In *2012 Computing in Cardiology*. 245–248.

[57] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, and others. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016), 484–489.

[58] Jenna Wiens, Eric Horvitz, and John V Guttag. 2012. Patient risk stratification for hospital-associated c. diff as a time-series classification task. In *Advances in Neural Information Processing Systems*. 467–475.

[59] B Williams, G Alberti, C Ball, D Bell, R Binks, L Durham, and others. 2012. National early warning score (NEWS): Standardising the assessment of acute-illness severity in the NHS. *London: The Royal College of Physicians* (2012).

[60] Xiaolin Yang, Seyoung Kim, and Eric P. Xing. 2009. Heterogeneous multitask learning with joint sparsity constraints. In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.). Curran Associates, Inc., 2151–2159. http://papers.nips.cc/paper/3706-heterogeneous-multitask-learning-with-joint-sparsity-constraints.pdf

[61] Safoora Yousefi, Congzheng Song, Nelson Nauata, and Lee Cooper. 2016. Learning Genomic Representations to Predict Clinical Outcomes in Cancer. *arXiv preprint arXiv:1609.08663* (2016).

[62] Jack E Zimmerman, Andrew A Kramer, Douglas S McNair, and Fern M Malila. 2006. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for todayâĂŹs critically ill patients. *Critical care medicine* 34, 5 (2006), 1297–1310.