Learning Tasks for Multitask Learning: Heterogenous Patient Populations in the ICU

Harini Suresh[*]
Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Laboratory
hsuresh@mit.edu

Jen J. Gong[*]
Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Laboratory
jengong@mit.edu

John V. Guttag
Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Laboratory
guttag@mit.edu

ABSTRACT

Machine learning approaches have been effective in predicting adverse outcomes in different clinical settings. These models are often developed and evaluated on datasets with heterogeneous patient populations. However, good predictive performance on the aggregate population does not imply good performance for specific groups.

In this work, we present a two-step framework to 1) learn relevant patient subgroups, and 2) predict an outcome for separate patient populations in a multi-task framework, where each population is a separate task. We demonstrate how to discover relevant groups in an unsupervised way with a sequence-to-sequence autoencoder. We show that using these groups in a multi-task framework leads to better predictive performance of in-hospital mortality both across groups and overall. We also highlight the need for more granular evaluation of performance when dealing with heterogeneous populations.

CCS CONCEPTS

• **Applied computing → Health informatics**; • **Computing methodologies → Multi-task learning**;

KEYWORDS

clinical risk models, multi-task learning, patient subpopulation discovery

ACM Reference Format:
Harini Suresh, Jen J. Gong, and John V. Guttag. 2018. Learning Tasks for Multitask Learning: Heterogenous Patient Populations in the ICU. In *Proceedings of Knowledge Discovery and Data Mining (KDD 2018).* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.1145/3219819.3219930

1 INTRODUCTION

Many important applications of machine learning utilize data from groups with different characteristics. Models trained on these datasets may not result in good predictions for each constituent group. This has been illustrated in tasks such as image classification [37], face recognition [3], and advertising [11]. In this work, we investigate this problem in clinical data, where such datasets are prevalent.

Machine learning models developed for clinical prediction tasks have the ability to aid care staff in deciding appropriate treatments. However, these clinical decision-making tools typically are not developed with specific subpopulations in mind, or they are developed for a single subpopulation and can suffer from data scarcity. The existence of these different subpopulations gives rise to a multifaceted problem: 1) a single model built for the entire patient population in aggregate does not imply equally good performance across distinct patient subpopulations, and 2) separate models learned on each of the distinct patient subpopulations do not take advantage of the shared knowledge that is common across patient subgroups.

Our solution combines *cohort discovery* with a *multi-task learning* model. Together, these steps form a pipeline that leverages shared information across distinct patient cohorts while accounting for their differences. During cohort discovery, we learn distinct patient subgroups in a data-driven way. These groups allow us to utilize a multi-task prediction framework where distinct patient groups are separate *tasks*. In order for multi-task learning to work effectively in this setup, examples need to be grouped into subpopulations that are sufficiently distinct with relation to the outcome of interest so that separate task models are beneficial.

Task formulations for multi-task learning with clinical data fall into two categories: 1) distinct *outcomes* are used as tasks [4, 18, 34, 42] and 2) distinct *patient populations* are used as tasks [30, 46]. Our formulation falls in the second category, where different *patient populations* are regarded as different tasks. Prior work has investigated pre-defined task definitions (e.g., [42]), and other work has used billing diagnosis codes to define latent bases for each patient [30]. In this work, we use physiological time-series dynamics to group examples into meaningful clinical tasks.

We investigate these methods in the context of building predictive models for patients in intensive care units (ICUs), using data from the publicly available MIMIC-III intensive care dataset [23].

Although patients in the ICU are typically more severely ill than patients in the hospital at large, the heterogeneity of patients in the ICU provides a useful case study for our approach, and MIMIC, as a publicly accessible dataset, enables reproducible studies.

We focus on the task of predicting whether a patient will die in the hospital, using data from the initial duration (24 hours or 48 hours) of their stay. Mortality prediction is an important task in clinical settings because a high risk of mortality is a signal of declining state and need for intervention. We show that a) there are salient subpopulations in the data that we can discover, and b) a multi-task model with subpopulations as tasks can outperform a

[*] The first two authors contributed equally to this work.

KDD 2018, August 2018, London, UK

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219930

single model that ignores subpopulation differences (a *global* model) as well as a single model trained on each subpopulation (*separate* models) on both overall and per-group performance metrics.

We also demonstrate the importance of performing granular evaluations across important subpopulations in a dataset. While much work reports overall metrics of performance, we highlight how this can hide underperformance on specific groups.

In Section 2, we describe the literature in machine learning for healthcare pertaining to 1) patient cohort discovery, and 2) multi-task learning. In Section 3, we describe the data we use. Next, we describe our two-step model formulation in Section 4, and our experiments and results in Sections 5 and 6.

2 RELATED WORK

The rapid adoption and availability of electronic health records (EHRs) has enabled new investigations into data-driven clinical support [14, 31, 44]. The broad goal of these studies is to learn from datasets of patient records in order to provide personalized treatment to patients. We provide a brief overview of work specifically in patient cohort discovery and multi-task learning.

2.1 Patient Cohort Discovery

Work in patient cohort discovery has focused on finding phenotypic characteristics of patients relevant for clinical insights, diagnoses, or risk-stratification. Constructing these groups requires finding a robust and meaningful representation of a patient's state.

2.1.1 Patient Representations. Static risk scores such as the Simplified Acute Physiology Score (SAPS II) [24] can be used to characterize a patient's state; these scores use a limited number of variables and do not take into account temporal trends [38]. Many recent works aim to *learn* data-driven representations of a patient's state. Some of these are learned in a supervised framework: for example, using the representation learned in a hidden layer of a deep neural network as a representation of patient state [5]. Other works characterize evolving patient state in an unsupervised way, inferring topics from clinical notes using Latent Dirichlet Allocation (LDA) [15], or inferring states and transitions with a switching state autoregressive model [16].

2.1.2 Cohort Discovery. After constructing a meaningful representation, *cohort discovery* requires using this representation to group patients into relevant cohorts. There is a broad range of what is considered a cohort (sometimes referred to as a *phenotype* in the literature) and how they are learned. In some cases, cohorts are pre-defined: for example, Gehrmann et al. have a group of physicians manually annotate examples with a set of 10 disease-related cohort classifications [13]. The process of manual annotation, however, is time-consuming, expensive and hard to scale. With the growing availability of large, high-dimensional clinical data, many works have proposed approaches to learning patient phenotypes [5, 20, 21, 33]. In all of these works, the patient cohorts are either analyzed for clinical insight, or used as additional features in a supervised prediction problem with a single, global model. In contrast to these works, we use the learned cohorts in a multi-task framework so that we can explicitly optimize for performance on each cohort.

2.2 Multi-task Learning for Clinical Risk-Stratification

The goal of multi-task learning is to combine learning of multiple related tasks, in order to improve performance across tasks (as opposed to learning each independently). Zhang and Yang present a comprehensive overview of multi-task methods [48], and Ruder give an overview of implementations of multi-task learning with deep neural networks [35].

In the clinical space, multi-task models have been used in a framework where the tasks are different prediction problems: for example, Harutyunyan et al. train a multi-task recurrent neural network that predicts mortality, length of stay, and ICD-9 groupings [18], Razavian et al. compare multi-task convolution and recurrent neural networks for predicting a number of ICD-9 diagnoses [34], and Choi et al. use recurrent neural network architecture to predict both diagnoses and the duration until the next visit [7]. Ngufor et al. use a multi-task model to improve prediction of various outcomes related to surgical procedures [29]. Wang et al. directly compare a multi-task model with many single-task models to demonstrate the utility of transferring knowledge across tasks for disease prediction [41]. Other work has explored post-learning strategies to cluster similar tasks in a multi-task model to enable optimal cross-transfer of knowledge [28]. Hierarchical models have also been used to predict multiple outcomes [36].

Predicting multiple outcomes aims to improve the generalizability of a model, whereas our goal is to build the best-suited model for distinct patient subpopulations by using the populations as the different tasks. Nori et al. do this by constructing a small number of latent basis tasks each with their own parameter vectors, and representing each patient as a combination of these tasks [30]. The specific combination is determined by the patient's record of diseases, represented as ICD-10 codes. Similarly, [46] uses a framework where patient-specific tasks are formulated as a linear combination of a shared set of base models. We consider salient and characterizable patient subpopulations, rather than separate tasks for each individual patient.

Other work aims to identify patient cohorts and transfer knowledge between them in a prediction framework. For example, hierarchical models have been used to take into account population differences [1, 10, 12]. Alaa et al. discover latent "classes" in the data, train Gaussian Processes to model the physiological data stream for each class, and transfer knowledge learned about the clinically stable population to a clinically declining population [1]. Our method has a similar aim (discovering groups in the data and utilizing shared knowledge across these groups) though we do not assume the framework of transferring knowledge from clinically stable to declining populations.

Our two-step pipeline enables us to learn patient subgroups that we use as tasks in a multi-task framework. In addition, it leverages the underlying physiological data of the patient rather than domain knowledge or auxiliary labels to discover relevant patient cohorts.

3 DATA

We use data from the publicly available MIMIC-III database [23]. Although MIMIC-III primarily contains data from a critical care setting, it has a large, heterogeneous patient population, and the

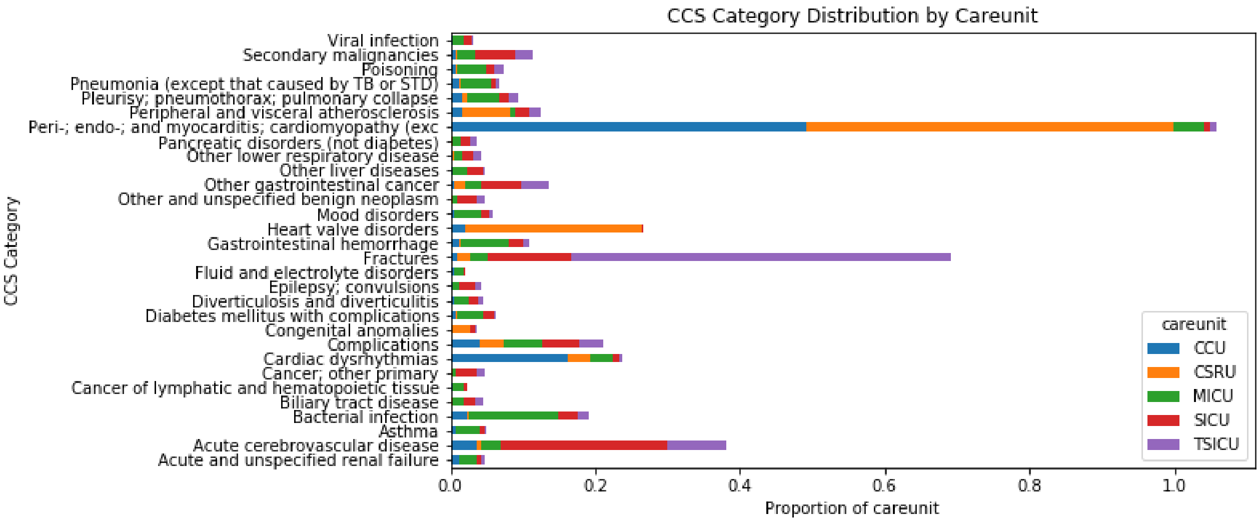


Figure 1: Primary diagnoses for patient admissions by Clinical Classifications Software (CCS) categories.

conclusions we draw from it in this work are likely relevant considerations for prediction tasks in other clinical settings. In addition, the dataset is made publicly available to researchers, enabling reproducibility. The dataset contains both structured electronic health record-like data, as well as free text clinical notes. We utilize the highly sampled vitals signs and irregularly sampled lab test results from the structured data, as well as static demographic attributes such as age, gender, and ethnicity. Table 2 contains a full list of features used in our experiments. Prior work has used these time-series to understanding patient physiological state to predict various outcomes such as intervention administration and mortality [5, 16, 39, 45].

Table 1: Number of adult patients and rate of in-hospital mortality (defined using the earliest time of mortality, or a note of "do not resuscitate" (DNR) or "comfort-measures only" (CMO) in each intensive care unit (ICU).

Careunit	*N*	*n*	Class Imbalance	Age (Mean)	Gender (Male)
CCU	4,905	339	0.069	82.56	0.58
CSRU	6,969	137	0.020	69.46	0.67
MICU	11,395	1118	0.098	77.98	0.51
SICU	5,178	397	0.077	72.57	0.52
TSICU	4,239	283	0.067	67.14	0.61
Overall	32,686	2274	0.070	74.59	0.57

Table 2: Physiological variables used for prediction.

Static Variables	Gender	Age	Ethnicity
Vitals and Labs	Anion gap	Bicarbonate	blood pH
	Blood urea nitrogen	Chloride	Creatinine
	Diastolic blood pressure	Fraction inspired oxygen	Glascow coma scale total
	Glucose	Heart rate	Hematocrit
	Hemoglobin	INR*	Lactate
	Magnesium	Mean blood pressure	Oxygen saturation
	Partial thromboplastin time	Phosphate	Platelets
	Potassium	Prothrombin time	Respiratory rate
	Sodium	Systolic blood pressure	Temperature
	Weight	White blood cell count	

* International normalized ratio of the prothrombin time

Patient characteristics are summarized in Table 1 and Figure 1. In particular, we note that the patients in different care units have very different rates of mortality, ranging from 2.0% in the Cardiac Surgery Recovery Unit (CSRU) to 9.8% in the Medical Intensive Care Unit (MICU). In addition, we note that patients in different units often present with different conditions, from acute events such as bone fractures to chronic conditions such as hypertension and coronary artery disease. Figure 1 shows the presence of some different disease categories.

4 METHODS

In this section, we describe our two-step procedure for 1) identifying meaningful patient cohorts, and 2) leveraging these cohorts as separate tasks in a multi-task learning framework. [1] This pipeline is diagrammed in Figure 2.

4.1 Identifying Meaningful Patient Cohorts

We utilize unsupervised representations and cohort-discovery methods for identifying relevant patient cohorts. Importantly, this method relies only on attributes at hospital admission or data from the initial portion of the patient's stay. Using this interval of data for patient phenotyping allows us to 1) identify patient phenotypes that are relevant when longitudinal patient history may not be immediately available, and 2) utilize only information prior to the time at which we make a prediction.

The raw patient data is a sparse timeseries; in order to discover cohorts we first encode this raw data into a dense fixed-length representation that we then cluster. We use a long short-term memory (LSTM) [22] autoencoder to produce a dense representation that captures important facets of the input. LSTMs have effectively modeled complicated dependencies in many types of time-series

[1] Model code is available at github.com/mit-ddig/multitask-patients.

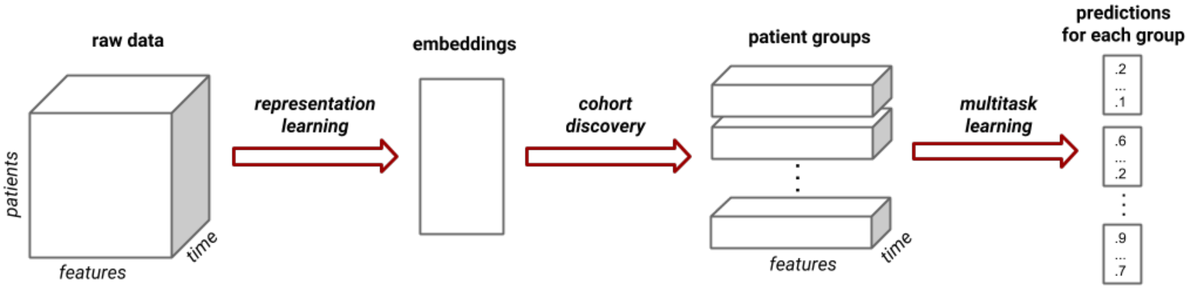
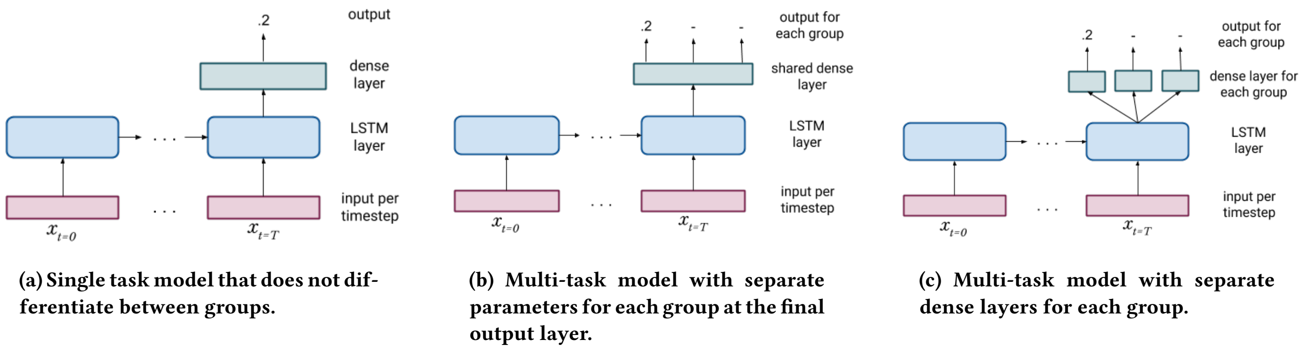


Figure 2: We present a two-step pipeline for 1) discovering relevant cohorts from the underlying physiological data for the prediction task at hand, and 2) using multi-task learning to share knowledge across related data while allowing distinct models to make predictions for different patient populations.



(a) Single task model that does not differentiate between groups.

(b) Multi-task model with separate parameters for each group at the final output layer.

(c) Multi-task model with separate dense layers for each group.

Figure 3: Single task and multi-task model configurations. Single task models have shared parameters for all examples, while multi-task models have separate parameters for each group in the output layer and/or the final dense layer.

data [2, 9, 19, 47], including clinical time-series [6, 25, 34, 39]. They are well-suited to our task because of the complex temporal dependencies in physiological time-series. Autoencoders have been used to learn compact representations of patient state from multi-modal timeseries EHR data [27, 40].

We use one LSTM layer in the encoder and one in the decoder. The embedded representation is the state of the encoder LSTM at the final timestep. This representation is then used in the decoder to reconstruct the input timeseries.

Embedding size was tuned for the reconstruction loss based on the training and validation data. Because reconstruction loss will consistently decrease when the embedding size is increased, we chose the embedding size based on the elbow in the reconstruction curve. We then cluster the embeddings with a Gaussian Mixture Model (GMM). The cluster assignments are used to group patients into tasks for the multi-task model.

4.2 Learning Predictive Models

In the prediction step, in order to go from a patient timeseries to a mortality prediction, we use an LSTM for all of the model configurations.

Our proposed approach uses a *multi-task* model, and we compare against several *single-task* baselines. The differences in these model configurations and training procedures are discussed in this section.

4.2.1 Single Task Model. The single task model (Figure 3a) consists of a single LSTM layer with a ReLU activation function followed by a single fully-connected layer with a sigmoid activation function. The output of the fully-connected layer is an estimate of the probability of mortality for the given example. We train this single task model on all the data to produce the global model baseline, and separately on data from each group to produce the separate model baselines.

4.2.2 Multi-task Model. In the multi-task model, our goal is to combine shared, global parameters along with separate parameters trained specifically for each group. In order to do this, we use the hard parameter-sharing framework of multi-task learning introduced in [4].

Like the single-task model, the multi-task model has one LSTM layer. The multi-task model was used either a single separate fully-connected layer for each group (Figure 3c) *or* a shared dense layer with separate weights leading to the output ((Figure 3b). During our grid search for model configurations, we limited the size of these fully-connected layers compared to the fully-connected layer of the single task model to ensure that both configurations were able to have similar capacity for making a fair comparison. The task-specific parameters are trained using only the losses from examples belonging to the task.

We compared our multi-task learning approach against two single-task approaches: 1) a separate single task model for each group, and 2) a global model for all patients, agnostic to task membership.

4.3 Evaluating Predictive Models Across Patient Cohorts

Machine learning models for clinical outcome predictions often utilize aggregate discriminative metrics such as the area under the receiver operating characteristic curve (AUC) to account for class imbalance (e.g., [5, 17, 39]). In settings where evaluations on specific patient cohorts is of interest, evaluation is more challenging. To evaluate metrics over different populations or outcomes, *micro* and *macro* versions of predictive metrics are used. In the *micro* case, all of the predicted probabilities for all patients are treated as if they come from a single classifier:

$$Metric_{micro} = Metric([\hat{y}_0, \cdots, \hat{y}_K], [y_0, \cdots, y_K]), \tag{1}$$

where K = the number of groups, $\hat{y}_k$ = predictions for the examples in group k and y_k = true labels for the examples in group k. This is the metric that is typically used in the literature. However, using these micro-evaluated metrics makes it difficult to assess how a model is performing on different subpopulations. This is especially true when the subpopulations are not equally represented.

Macro measures evaluate a metric *within* each cohort first, and then average the results across cohorts:

$$Metric_{macro} = \frac{1}{K}\sum_{k=0}^{K} Metric(\hat{y}_k, y_k) \tag{2}$$

This metric is better suited to assess performance across groups of disparate size, since each group contributes equally to the macro metric evaluation [26].

We use both of these methods of computing metrics, and evaluate micro- and macro- AUC. We additionally evaluate micro- and macro- positive predictive value (PPV) and specificity at a sensitivity of 80%. While AUC gives a sense of overall discriminative model performance, we show PPV and specificity at a single decision threshold to evaluate how well such a model might perform in a real setting.

5 EXPERIMENTS

We developed models for predicting in-hospital mortality using physiological time-series data from the initial portion of the patient's ICU stay.

5.1 Prediction Task Definition

We define *in-hospital mortality* as having an outcome of mortality, *or* a note of "Do Not Resuscitate" (DNR) or "Comfort Measures Only" (CMO). This definition is in contrast to what has been used in prior work, where only mortality was considered (e.g., [15, 17, 18]). Notes of DNR or CMO indicate differences in what clinical interventions will be taken, and our proposed risk models might not have actionable predictions.

We conducted experiments in two settings:

(1) Using the first 24 hours of data from the patient's stay to predict in-hospital mortality starting at 36 hours into the stay. Acuity scores such as the Simplified Acute Physiology Score (SAPS-II) [24] also use the first 24 hours of data to evaluate patient severity of illness.
(2) Using the first 48 hours of data from the patient's stay to predict in-hospital mortality starting 72 hours in to the stay. We explore this task because the first 24 hours often contain routine tests done upon admission, and this time period might reflect different changes in patient physiology.

Each of the described experiments includes prediction gaps between the information used about a patient and the point at which outcomes are counted. This is common in the literature, and the motivation is two-fold: 1) it eliminates trivial cases where the outcome is imminent, and 2) simulates a situation in which there is time to intervene. Patients who were discharged or had an outcome of in-hospital mortality during the period of the stay being used for prediction or during the gap period were dropped from the experiments.

5.2 Data Processing

We considered all ICU patients over the age of 15 and took the patient's first ICU stay (if there are multiple), as the majority of patients have only one ICU stay. For each patient, we extracted 29 time-varying vitals and labs, detailed in Table 2. The timestamps of these measurements were rounded to the nearest hour. If an hour had multiple measurements for a signal, those measurements were averaged. We created discrete, binary features by first transforming each variable to the z-score, and then making each z-score value its own column. Similar methods have been used in previous work [40, 45] in order to have an explicit representation of missing values, and to avoid overfitting to small changes in the physiological variables. This creates a very sparse data representation.

In addition, we include demographics such as the patient's ethnicity, gender and age quartile. These static variables are replicated across all time-steps for a patient.

5.3 Model Implementation and Training

In this section, we describe our model training and selection procedures. We describe 1) the supervised models trained for predicting the outcome using a single-task and multi-task framework, 2) the autoencoders used to learn unsupervised, latent representations for our physiological data, and 3) the Gaussian Mixture Models used to identify cohorts from the latent representations of the data. For all experiments, we split the data using an 80:20 training:test split stratified on the outcome.

5.3.1 Gaussian Mixture Model. We used the Scikit-learn (version 0.19.1) implementation of Gaussian Mixture Models [32]. The GMM was initialized using assignments from k-means clustering. We fit the model with 30 different initializations and chose the model that gave the highest data likelihood. We divided the training data in a 7:1 training:validation split. We explored several possible values for the number of clusters, and chose the value that resulted in the best predictive performance on the validation set when the clusters were used as tasks in the multi-task model.

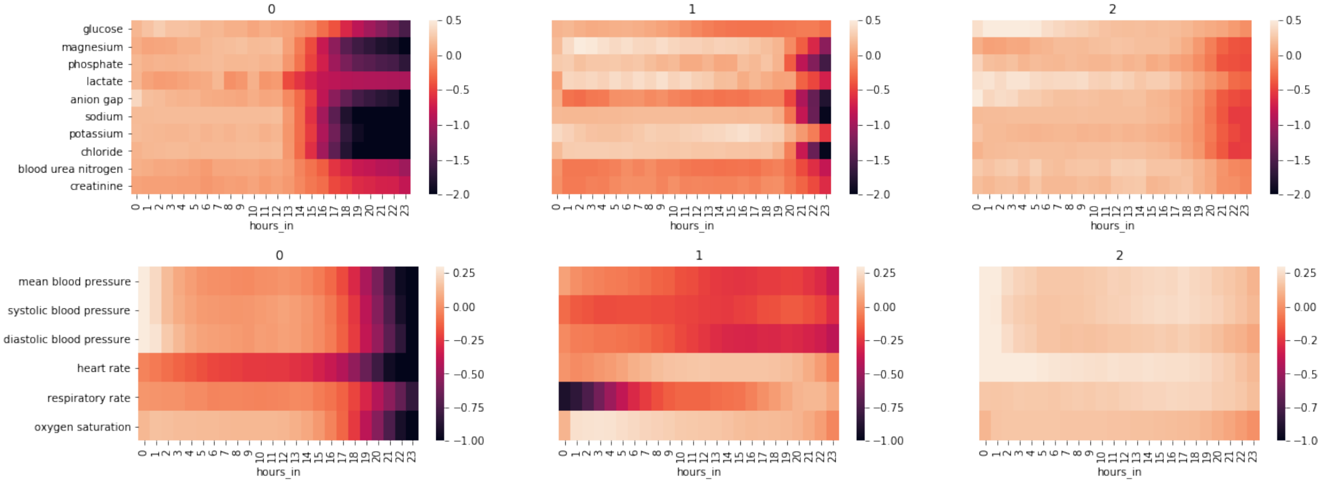


Figure 4: Selected lab test and vital signs features over the first 24 hours for the unsupervised clusters. Figures show z-score values of the features over time; 0 indicates the mean value, positive values indicate elevated measures, and negative values indicate decreased measures. In the first 24 hours, lab test results in cluster 2 are more elevated than in cluster 0 and cluster 1. Cluster 0 has a centroid with decreasing heart rate, whereas cluster 1's centroid shows an increasing heart rate. Note that while the trends are opposing, both centroids have heart rate values that are below the mean. Additionally, we note that blood test results for magnesium, lactate, and potassium are elevated in cluster 1, while glucose is elevated in cluster 2.

5.3.2 Unsupervised Representations. To learn unsupervised representations, we used a sequence-to-sequence autoencoder with LSTM units implemented with Keras. We explored several hidden dimension sizes for the autoencoder, and chose the dimensionality corresponding to the elbow in the reconstruction error curve on the validation set. This procedure resulted in an embedding size of 100.

The autoencoder was trained with a mean squared error loss function and the Adam optimizer with an initial learning rate of 0.001. We trained the autoencoder for a maximum of 100 epochs; to prevent over-training, we employed early stopping if the validation loss decreased for 6 epochs.

5.3.3 Single and Multi-task Prediction Models. We implemented the single and multi-task models using Keras version 2.1.3 [8]. We determined the best model configurations by doing a grid search over possible hyperparameters and choosing the best configuration over 5 random splits of the training data into 7:1 training:validation splits. We allowed the global model to search over a larger range of layer sizes to enable a fair comparison with the extra parameters that could be introduced in the multi-task model. We used binary cross-entropy as our loss function, and the Adam optimizer with a learning rate of 0.0001. The models were trained for a maximum of 100 epochs with early stopping.

6 RESULTS

We report results comparing the global single-task model with the multi-task model. We also tested a baseline of using separate single-task models for each task, but this model had significantly worse performance in all cases so we have not included it. All reported statistical significance results were computed using the Wilcoxon signed-rank test [43] over 100 bootstrapped samples of the test set.

Table 3: Cohort statistics at 24 hours and 48 hours

	Cohort Type	Group	N	n	Class Imbalance
24 hours	**Unsupervised**	0	11862	404	0.0341
		1	6434	107	0.0166
		2	14390	1786	0.1241
	Global	-	32686	2297	0.0703
48 hours	**Unsupervised**	0	13433	291	0.0217
		1	16995	1436	0.0845
	Global	-	30,428	1,727	0.0568

Bootstrapped samples were of the same size and class imbalance as the original test set.

6.1 Predicting Mortality at 24 Hours

6.1.1 Discovered Cohorts are Physiologically Distinct. Statistics about the discovered cohorts are shown in Table 3, and Figure 4 shows visualizations of the tasks learned using our methodology.

Table 4 shows that the three cohorts of patients discovered from the first 24 hours of data are different in terms of size and class imbalance. While two of the clusters are large, with over 10,000 patients each, the class imbalances in these two cohorts are dramatically different. Whereas Cohort 0 has an outcome incidence of 3%, Cohort 1 has an outcome incidence of 12%.

In addition, the centroids of the cohorts show physiological trends over the first 24 hours that differ in important ways (Figure 4). For example, clusters 0 and 2 both have elevated blood pressure in the first several hours of their stay. However, whereas cluster 0's blood pressure decreases over time, cluster 2's blood

Table 4: 24 Hour Mortality Prediction: Performance differences between multi-task and global models on specific cohorts. A multi-task model with pre-defined tasks based on careunits performs poorly, while the unsupervised multi-task model performs comparably on two out of three cohorts and better on one. Significant differences ($p < 0.01$) are shown in bold.

		AUC		PPV		Specificity	
Cohort type	Cohort	Global	Multi-task	Global	Multi-task	Global	Multi-task
Unsupervised	0	0.803	**0.819**†	0.083	**0.103**‡	0.732	**0.786**‡
	1	0.811	**0.829**†	0.120	**0.126***	0.916	0.915
	2	0.814	**0.821**‡	0.276	**0.288**‡	0.734	**0.742**‡
	Macro	0.809	**0.823**†	0.159	**0.172**‡	0.794	**0.814**‡
	Micro	0.852	**0.858**‡	**0.231**$^{\diamond}$	0.228	**0.817**†	0.814
Careunits	CCU	**0.862***	0.861	**0.248**‡	0.229	**0.834**‡	0.819
	CSRU	0.849	**0.867**†	0.107	**0.117**†	0.893	**0.898**†
	MICU	0.814	**0.832**‡	0.261	**0.262***	0.764	**0.766***
	SICU	0.839	**0.855**†	0.226	**0.238**†	0.781	**0.796**†
	TSICU	0.846	**0.869**‡	0.183	**0.192**†	0.823	**0.818**$^{\diamond}$
	Macro	0.842	**0.857**‡	0.205	**0.208**†	0.819	0.819
	Micro	0.852	**0.866**‡	0.231	**0.233**$^{\diamond}$	0.817	**0.821**†

$\star$: $0.01 > p > 0.001$, $\diamond$: $0.001 > p >$ 1e-5, $\dagger$: 1e-5 $> p >$ 1e-15, $\ddagger$: $p <$ 1e-15

pressure stays elevated. We also observe that the heart rate in cluster 0 decreases over time, whereas cluster 1 and cluster 2 both have increasing heart rates. The differences between these cluster centroids indicate that our method of learning dense representations from the sparse physiological data for clustering discovers salient differences between patients.

6.1.2 Multi-task Models Outperform Global and Separate Models. Our multi-task framework significantly improved performance over the global model in AUC, PPV, and specificity on each of the learned cohorts (p < 0.01). In addition, it improved performance on aggregate metrics such as macro-AUC, PPV, and specificity, as well as micro-AUC. However, micro- PPV and micro-specificity were significantly worse. This is because micro-metrics are computed by setting a *single* threshold across all examples, regardless of the cohort they belong to. However, setting a single threshold ignores the large class imbalance differences between the cohorts. In contrast, the macro-measure, which considers a separate decision threshold based on 80% sensitivity for each individual cohort, is significantly better when using the multi-task model compared to the global model. The performance increase from using the multi-task model indicates that we can improve both per-group and aggregate measures using this framework.

More generally, we hope to highlight the importance of evaluating methods across subpopulations, since overall micro-measures can hide underperformance on specific subgroups. For example, the global model achieves an overall Micro AUC of 0.852, but it's AUC on cluster 0 was only 0.803. Without an evaluation broken down by groups, it would be hard to detect such performance disparities.

We contrast our learned patient populations against expert knowledge driven cohorts, where patients are stratified by the first care unit they are admitted to. This cohort definition does not rely on the underlying physiological data. However, it is a reasonable attribute on which to split patients, given the differences across care units in patient conditions (see Figure 1). In addition, the rate of adverse events in these different units is highly variable, from less than two percent in the Cardiac Surgery Recovery unit to over 10% in the Medical ICU.

Grouping patients by first care unit and using these groups as tasks in an multi-task framework significantly improved performance over the global model. At this point in the patient's stay, first care unit is likely a meaningful indicator of differences between populations. However, while we have access to meaningful patient cohorts defined by first care unit, such distinct, labeled groups may not be available for a different clinical population. Our unsupervised method results in significant improvements, without requiring expert knowledge.

6.2 Predicting Mortality at 48 Hours

In contrast to the results from predicting mortality at 24 hours, our multi-task model with learned patient cohorts does not result in significant improvements compared to the global model when predicting mortality after 72 hours using 48 hours of data. One reason for this may be the sparse nature of the physiological data. Because routine lab tests and other evaluations are frequently done in the first day of a patient's ICU stay, data presence drops off in the second day. Because of this, the data are heavily biased towards missing values; therefore, the autoencoder we use to construct dense representations of patient physiological state may also be biased.

The macro- and micro- performance metrics are shown in Table 5. For our unsupervised method, macro AUC and PPV were not significantly different from the global model's performance, but the specificity was significantly worse. In addition, we again compared a multi-task model with learned cohorts against the expert-defined cohorts. In this case, while our method did not result in significant differences, the care units multi-task model performed significantly worse on all metrics compared to the global model. As a patient's stay in the ICU progresses, her characteristics may be less defined

Table 5: 48 Hour Mortality Prediction: Performance differences between multi-task and global models on specific cohorts. A multi-task model with pre-defined tasks based on careunits performs poorly, while the unsupervised multi-task model performs comparably. Significant differences ($p < 0.01$) are shown in bold.

		AUC		PPV		Specificity	
Cohort type	Cohort	Global	Multi-task	Global	Multi-task	Global	Multi-task
Careunits	Macro	**0.859**‡	0.839	**0.187**‡	0.170	**0.833**‡	0.826
	Micro	**0.865**‡	0.856	**0.206**†	0.198	**0.833***	0.832
Unsupervised	Macro	0.834	0.833	0.154	0.154	**0.789**‡	0.775
	Micro	**0.865**†	0.861	0.206	0.191	**0.833**‡	0.812

$\star$: 0.01 > p > 0.001, $\diamond$: 0.001 > p > 1e-5, $\dagger$: 1e-5 > p > 1e-15, $\ddagger$: $p < 1e - 15$

by the care unit she is admitted to compared to the interventions that are being administered. This highlights the need to use the underlying data to discover meaningful and distinct cohorts as tasks, and motivates further research on how to discover such cohorts in the presence of extreme sparsity (as in the 48-hour data).

7 CONCLUSIONS & DISCUSSION

In this work, we show how machine learning models trained globally on heterogeneous populations can perform well in an overall sense while under-performing on specific, meaningful populations. We propose a two-step pipeline that 1) identifies distinct patient subpopulations, and 2) leverages these subpopulations in a multi-task framework to effectively share knowledge.

We demonstrate that for 24-hour mortality prediction, our learned cohorts significantly improve over a single model learned on all of the data. In addition, we compare against an expert-knowledge driven method for identifying cohorts. We show that meaningful, distinct tasks can be learned in a data-driven way without pre-specifying cohorts for a particular outcome. We evaluate our models on the overall population, and on each separate cohort.

We highlight the need to evaluate performance across relevant cohorts. Much real data consists of heterogeneous populations, and reporting a single, overall evaluation metric can hide disparities in performance across groups. Accounting for these patient differences is important in model training, but also in model *evaluation*.

In addition, we believe that while unsupervised clustering of the physiological data representations led to improved results in the multi-task framework, learning clusters and representations that are guided by the specific outcome of interest could lead to useful outcome-specific cohorts. While unsupervised cohorts are generalizable across outcomes, representations and cohorts that are outcome-specific could lead to further improvements in predictive performance. For example, patient subpopulations that are distinct for predicting ventilator administration may look very different compared to patient subpopulations that are distinct for predicting length of stay or discharge status.

While the work we present is specific to the MIMIC-III dataset, we believe that the considerations we outline here are broadly applicable to clinical prediction tasks. We hope the ideas we have discussed can help ensure that machine learning algorithms are not assumed to be one-size-fit-all, but rather that they work well for all groups involved.

8 ACKNOWLEDGMENTS

This research was funded in part by the National Institutes of Health under award number P50DC015446 and by Wistron Corporation.

REFERENCES

[1] Ahmed M Alaa, Jinsung Yoon, Scott Hu, and Mihaela van der Schaar. 2016. Personalized risk scoring for critical care patients using mixtures of Gaussian Process Experts. *arXiv preprint arXiv:1605.00959* (2016).

[2] 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.

[3] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[4] Rich Caruana. 1993. Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *ICML*.

[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, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*. 301–318.

[8] François Chollet et al. 2015. Keras. https://github.com/keras-team/keras. (2015).

[9] 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.

[10] Marine Corbin, Lorenzo Richiardi, Roel Vermeulen, Hans Kromhout, Franco Merletti, Susan Peters, Lorenzo Simonato, Kyle Steenland, Neil Pearce, and Milena Maule. 2012. Hierarchical regression for multiple comparisons in a case-control study of occupational risks for lung cancer. *PloS one* 7, 6 (2012), e38944.

[11] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112.

[12] Paola D'Errigo, Maria E Tosti, Danilo Fusco, Carlo A Perucci, and Fulvia Seccareccia. 2007. Use of hierarchical models to evaluate performance of cardiac surgery centres in the Italian CABG outcome study. *BMC medical research methodology* 7, 1 (2007), 29.

[13] Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T Carlson, Joy T Wu, Jonathan Welt, John Foote Jr, Edward T Moseley, David W Grant, Patrick D Tyler, et al. 2017. Comparing Rule-Based and Deep Learning Models for Patient Phenotyping. *arXiv preprint arXiv:1703.08705* (2017).

[14] Marzyeh Ghassemi, Leo Anthony Celi, and David J Stone. 2015. State of the art review: the data revolution in critical care. *Critical Care* 19, 1 (2015), 118.

[15] 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.

[16] Marzyeh Ghassemi, Mike Wu, Michael C Hughes, Peter Szolovits, and Finale Doshi-Velez. 2017. Predicting intervention onset in the ICU with switching state space models. *AMIA Summits on Translational Science Proceedings* (2017), 82.

[17] Jen J Gong, Tristan Naumann, Peter Szolovits, and John V Guttag. 2017. Predicting Clinical Outcomes Across Changing Electronic Health Record Systems. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1497–1505.

[18] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, and Aram Galstyan. 2017. Multitask Learning and Benchmarking with Clinical Time Series Data. *arXiv preprint arXiv:1703.07771* (2017).

[19] 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.

[20] Joyce C Ho, Joydeep Ghosh, Steve R Steinhubl, Walter F Stewart, Joshua C Denny, Bradley A Malin, and Jimeng Sun. 2014. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of biomedical informatics* 52 (2014), 199–211.

[21] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. 2014. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 115–124.

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

[23] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, 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), 160035.

[24] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. 1993. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *Jama* 270, 24 (1993), 2957–2963.

[25] 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).

[26] Christopher Manning, Prabhakar Raghavan, and Hinrich SchÃijtze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[27] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports* 6 (2016), 26094.

[28] C. Ngufor, S. Upadhyaya, D. Murphree, D. Kor, and J. Pathak. 2015. Multi-task learning with selective cross-task transfer for predicting bleeding and other important patient outcomes. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 1–8. https://doi.org/10.1109/DSAA.2015.7344836

[29] Che Ngufor, Sudhindra Upadhyaya, Dennis Murphree, Nageswar Madde, Daryl Kor, and Jyotishman Pathak. 2015. A heterogeneous multi-task learning for predicting RBC transfusion and perioperative outcomes. In *Conference on Artificial Intelligence in Medicine in Europe*. Springer, 287–297.

[30] Nozomi Nori, Hisashi Kashima, Kazuto Yamashita, Susumu Kunisawa, and Yuichi Imanaka. 2017. Learning Implicit Tasks for Patient-Specific Risk Modeling in ICU.. In *AAAI*. 1481–1487.

[31] Ziad Obermeyer and Ezekiel J Emanuel. 2016. Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. *The New England journal of medicine* 375, 13 (2016), 1216.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[33] Rimma Pivovarov, Adler J Perotte, Edouard Grave, John Angiolillo, Chris H Wiggins, and Noémie Elhadad. 2015. Learning probabilistic phenotypes from heterogeneous EHR data. *Journal of biomedical informatics* 58 (2015), 156–165.

[34] Narges Razavian, Jake Marcus, and David Sontag. 2016. Multi-task prediction of disease onsets from longitudinal laboratory tests. In *Machine Learning for Healthcare Conference*. 73–100.

[35] Sebastian Ruder. 2017. An Overview of Multi-Task Learning in Deep Neural Networks. *CoRR* abs/1706.05098 (2017). arXiv:1706.05098 http://arxiv.org/abs/1706.05098

[36] Trevor R Shaddox, Patrick B Ryan, Martijn J Schuemie, David Madigan, and Marc A Suchard. 2016. Hierarchical models for multiple, rare outcomes using massive observational healthcare databases. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 9, 4 (2016), 260–268.

[37] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. In *NIPS 2017 workshop: Machine Learning for the Developing World*.

[38] Tasnim Sinuff, Neill KJ Adhikari, Deborah J Cook, Holger J Schünemann, Lauren E Griffith, Graeme Rocker, and Stephen D Walter. 2006. Mortality predictions in the intensive care unit: comparing physicians with scoring systems. *Critical care medicine* 34, 3 (2006), 878–885.

[39] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017. Clinical Intervention Prediction and Understanding using Deep Networks. *Machine Learning for Health* (2017).

[40] Harini Suresh, Peter Szolovits, and Marzyeh Ghassemi. 2017. The use of autoencoders for discovering patient phenotypes. *arXiv preprint arXiv:1703.07004* (2017).

[41] Xiang Wang, Fei Wang, Jianying Hu, and Robert Sorrentino. 2014. Exploring joint disease risk prediction. In *AMIA Annual Symposium Proceedings*, Vol. 2014. American Medical Informatics Association, 1180.

[42] Jenna Wiens, John Guttag, and Eric Horvitz. 2016. Patient risk stratification with time-varying parameters: a multitask learning approach. *The Journal of Machine Learning Research* 17, 1 (2016), 2797–2819.

[43] Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics* 1, 6 (1945), 80–83.

[44] Jionglin Wu, Jason Roy, and Walter F Stewart. 2010. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical care* 48, 6 (2010), S106–S113.

[45] Mike Wu, Marzyeh Ghassemi, Mengling Feng, Leo A Celi, Peter Szolovits, and Finale Doshi-Velez. 2017. Understanding vasopressor intervention and weaning: Risk prediction in a public heterogeneous clinical time series database. *Journal of the American Medical Informatics Association* 24, 3 (2017), 488–495.

[46] Jianpeng Xu, Jiayu Zhou, and Pang-Ning Tan. 2015. Formula: F act OR ized MU lti-task L e A rning for task discovery in personalized medical models. In *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, 496–504.

[47] 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.

[48] Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114* (2017).