

AdaCare: Explainable Clinical Health Status Representation Learning via Scale-Adaptive Feature Extraction and Recalibration

Liantao Ma^{1,3}, Junyi Gao^{1,2}, Yasha Wang^{1,2*}, Chaohe Zhang^{1,3}, Jiangtao Wang⁴,
Wenjie Ruan⁴, Wen Tang⁵, Xin Gao^{1,3}, Xinyu Ma^{1,3}

¹Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing, China

²National Engineering Research Center of Software Engineering, Peking University, Beijing, China

³School of Electronics Engineering and Computer Science, Peking University, Beijing, China

⁴School of Computing and Communications, Lancaster University, UK

⁵Division of Nephrology, Peking University Third Hospital, Beijing, China

{malt, wangyasha}@pku.edu.cn, {jiangtao.wang, wenjie.ruan}@lancaster.ac.uk, tanggwen@126.com

Abstract

Deep learning-based health status representation learning and clinical prediction have raised much research interest in recent years. Existing models have shown superior performance, but there are still several major issues that have not been fully taken into consideration. First, the historical variation pattern of the biomarker in diverse time scales plays a vital role in indicating the health status, but it has not been explicitly extracted by existing works. Second, key factors that strongly indicate the health risk are different among patients. It is still challenging to adaptively make use of the features for patients in diverse conditions. Third, using prediction models as the black box will limit the reliability in clinical practice. However, none of the existing works can provide satisfying interpretability and meanwhile achieve high prediction performance. In this work, we develop a general health status representation learning model, named AdaCare. It can capture the long and short-term variations of biomarkers as clinical features to depict the health status in multiple time scales. It also models the correlation between clinical features to enhance the ones which strongly indicate the health status and thus can maintain a state-of-the-art performance in terms of prediction accuracy while providing qualitative interpretability. We conduct a health risk prediction experiment on two real-world datasets. Experiment results indicate that AdaCare outperforms state-of-the-art approaches and provides effective interpretability, which is verifiable by clinical experts.

Introduction

Health status prediction (e.g., mortality risk prediction, disease prediction) is of great interest to physicians. For inpatients or patients with chronic diseases who face severe life threats and receive long-term treatments, their health conditions are complex and continually changing over time. By predicting patient's health status, physicians can select personalized follow-up treatments, prevent adverse outcomes, assign medical resources effectively, and reduce the medical cost. Normally, some biomarkers, such as blood albumin

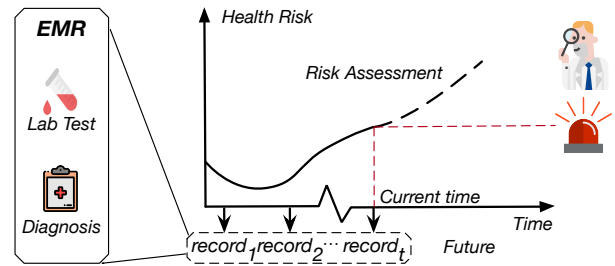


Figure 1: Perform dynamic patient health risk prediction to trigger the alarm for adverse outcomes and take an early individualized intervention.

and blood glucose, are recorded through the treatment trajectories and further have been taken into consideration for the prediction. In a practical diagnosis process, physicians need to comprehensively evaluate the health of patients by identifying the high-risk factors. The precise risk prediction requires a high level of clinical expertise and experience.

Nowadays, electronic healthcare information systems are widely used in various healthcare institutions, which can precisely record the lab test results and health information of patients in terms of Electronic Medical Records (EMR). As depicted by Figure 1, EMR can be seen as a type of multivariate time series data and provide essential healthcare information for the data-driven healthcare prediction.

Recently, due to the remarkable representation learning ability of deep neural networks, many deep learning-based models have been developed to tackle such prediction tasks by using EMR data, including mortality prediction (Ma et al. 2020), disease diagnosis prediction (Lee et al. 2018), and patient phenotype identification (Baytas et al. 2017). Usually, those models first embed the EMR data into low-dimensional feature space to learn the dense representation of the patients' health status and then perform specific clinical analysis tasks based on such representation. However, there are still some issues that are not yet fully resolved by existing research works, i.e., how to effectively embed temporal health information comprehensively, and how to assure the trustworthiness of the representation learning model

*Corresponding Author

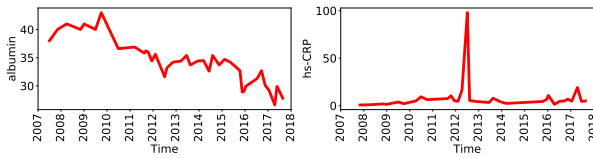


Figure 2: The long-term descending trend of blood albumin (left) and short-term abnormal rising of hypersensitive C-reactive protein (right) indicate the health risk.

in terms of providing verifiable interpretations. The issues are summarized as follows:

- **I_1 : Historical variation pattern of the biomarker in different time scales.** Not only the value of lab test results but also its variation pattern contains essential information on the health status. The long-term variation trend and short-term abnormal variation of different biomarkers both show the health status of patients from different aspects. Thus, it is vital that the variation of different biomarkers can be captured at different time scales. For example, for end-stage renal disease (ESRD) patients, the long-term descending trend of blood albumin is a strong indicator of malnutrition and health deterioration (Bharadwaj et al. 2016). On the contrary, the short-term abnormal rising of hypersensitive C-reactive protein indicates the high-risk clinical event (e.g., peritonitis) (Ma et al. 2018b). Although several research works try to utilize the convolutional operator to extract temporal patterns of clinical events (Cheng et al. 2016; Ma, Xiao, and Wang 2018), none of them can capture such patterns in multi-level time scales simultaneously.
- **I_2 : Adaptively making use of the clinical features for patients in diverse conditions.** Key factors that strongly indicate the health risk are different among patients (Valko and Hauskrecht 2010). So it is critical that the model can utilize the features adaptively when learning the health status representation and performing prediction for patients in different conditions. The model adaptability can be specifically expressed in two aspects:
 - The importance of features in different time-scales varies among different patients. For example, for the patient suffering from chronic disease, the feature extracted in the long term may be more representative for depicting the health status. On the contrary, for the patient diagnosed with acute disease, the short-term feature describes the health risk more precisely.
 - The importance of features should also be adaptive to different characteristics. For example, the model should pay more attention to creatinine and urea when they rise ¹, and the model should concentrate more on diastolic blood pressure when the patient has been diagnosed with cerebrovascular disease².

¹Plasma concentrations of creatinine and urea are usually associated with systemic manifestations (uremia) for chronic kidney disease patients (Anna Malkina 2018).

²The level of diastolic blood pressure is usually associated with

- **I_3 : Interpretability for various patients.** Medical experts need to understand how a certain decision is made by a model to a particular patient at different visits. So that the prediction results are trustworthy for developing individualized intervention and extracting medical knowledge (Tangri, Ansell, and Naimark 2011). For example, if the model triggers a health warning by taking the creatinine and urea as key factors, the physician is alerted to assess such patients for possible systemic manifestations. Moreover, it will also remind the physicians of the previously unknown correlation between the biomarker and death reason. However, most of the existing works can only provide visit-level or disease-feature-level interpretability by attention mechanism (Ma et al. 2017; Bai et al. 2018). As far as we know, RETAIN (Choi et al. 2016) is the only work that can provide reasonable biomarker-feature-level interpretability by utilizing two-level attention as an end-to-end model, but its prediction accuracy is unsatisfactory (Ma, Xiao, and Wang 2018; Ma et al. 2018a). To bridge this research gap, our model AdaCare can provide a fine-grained feature-level interpretation for the model prediction, but also achieves a state-of-the-art prediction accuracy.

By jointly considering the above research issues in clinical practice, we propose a clinical health status representation learning model via scale-adaptive feature extraction and recalibration, *AdaCare*³. It monitors biomarkers in long and short time scales simultaneously to extract temporal variation patterns, depicting the health status comprehensively for patients in diverse conditions (e.g., diagnosed with chronic diseases or acute diseases). *AdaCare* models the high relationship between clinical features to abstract the input. At each visit, *AdaCare* selects the most indicative medical features to build health status representation. Empirical studies show *AdaCare* boosts the performance and meanwhile offers key features that lead to the prediction. Our main contributions are summarized as follows:

- We build a general health status representation learning model, *AdaCare*, to effectively embed the health status and provide reasonable interpretability for patients in diverse conditions. *AdaCare* uses the dilated convolution with multi-scale receptive fields to capture the long and short-term variation patterns of biomarkers as clinical features and depict patient health status more comprehensively (addressing I_1).
- We build the scale-adaptive feature recalibration module, which explicitly and adaptively models the feature relationship based on squeeze-and-excitation block (Hu, Shen, and Sun 2018) to selectively enhances high-risk features and meanwhile suppress the useless ones (addressing I_2). And thus, *AdaCare* can provide interpretability on health status representation learning for patients in diverse health conditions as an end-to-end model, and further remind the physicians with the precursor of

cerebrovascular disease (Rabkin, Mathewson, and Tate 1978)

³We release our code and case studies at GitHub <https://github.com/Accountable-Machine-Intelligence/AdaCare>

health risk (addressing I_3). Such interpretability is indicative of understanding how the model utilizes EMR data to make the assessment and extract valuable medical knowledge.

- We conduct two prediction tasks (i.e., decompensation prediction and mortality prediction) on two real-world datasets (i.e., MIMIC-III dataset and end-stage renal disease dataset) respectively to verify the performance. The results show that AdaCare outperforms the baseline approaches in both tasks. The interpretability of AdaCare is demonstrated by an overall observation of feature recalibration. Besides, the obtained medical knowledge has been positively confirmed by clinical experts.

Related Work

Over the past ten years, there has been a massive explosion in the amount of digital information stored in electronic medical records, which opens a door for researchers to make secondary use of these records for various clinical applications. Deep learning-based models have shown the capability to perform mortality prediction, patients subtyping, and diagnosis prediction. Though the medical tasks vary from each other, their essences are usually learning the health status representations of patients. There are two essential concerns among deep-learning-based EMR analysis researches:

Temporal Medical Feature Extraction

Some researches tried to extract high-level temporal clinical features by convolution modules as well as performing healthcare prediction (Cheng et al. 2016; Ma, Xiao, and Wang 2018). However, according to medical experience, the variation of biomarkers should be evaluated in different time scales simultaneously when evaluating the patient’s condition. To the best of our knowledge, there has not been any research extracting clinical features in multiple time scales effectively.

Interpretability of EMR Analysis

On the one hand, the interpretability shown in most of the existing works mainly focuses on visit-level attention. For example, some researches proposed RNN-based models with attention mechanisms to measure the relationships of different visits (Ma et al. 2017; Lee et al. 2018).

On the other hand, several researches have also explored the interpretability in the medical-feature-level. Timeline (Bai et al. 2018) utilizes self-attention to generate clinical visit embedding, but can only identify disease code-level importance. Some researches show the importance of features via adversarial attack, which is not an end-to-end framework (Sun et al. 2018). RETAIN (Choi et al. 2016) is more closely related to our work in terms of interpretability, which achieves feature-level interpretability by using attention mechanisms. However, the prediction performance of RETAIN is limited (Ma, Xiao, and Wang 2018; Ma et al. 2018a), due to the deficiency of effective high-level clinical feature extraction. Existing studies still cannot capture the importance of biomarkers dynamically and

meanwhile gain a performance boost in an end-to-end deep learning-based healthcare predictive model.

Preliminary

A Motivating Example

We take the health status prediction of end-stage renal disease (ESRD) patients as the motivating example. Currently, many people are suffered from ESRD in the world (Tangri, Ansell, and Naimark 2011; Isakova et al. 2011). They face severe life threats and need lifelong treatments with periodic visits to the hospitals for multifarious tests (e.g., blood routine examination). The whole procedure needs a dynamic patient health risk prediction to help patients recover smoothly and prevent the adverse outcome, based on the medical records collected along with the visits. The core task of AdaCare is to learn the health status representation of the patient and perform the healthcare prediction.

Table 1: Notations used in AdaCare

Notation	Definition
y_t	Groundtruth of prediction target at t -th visit
\hat{y}_t	Prediction result at t -th visit
$\mathbf{r}_t \in \mathbb{R}^{N_r}$	Multivariate visit record at t -th visit
$\mathbf{u}_t^r \in \mathbb{R}^{ \mathbf{r}_t }$	Feature recalibration weight of \mathbf{r}_t
$\tilde{\mathbf{r}}_t \in \mathbb{R}^{ \mathbf{r}_t }$	Weighted input record
$\mathbf{c}_t \in \mathbb{R}^{K * N_c}$	Extracted convolutional embedding of $\mathbf{r}_{t-L+1:t}$
$\mathbf{u}_t^c \in \mathbb{R}^{ \mathbf{c}_t }$	Scale-adaptive recalibration weight of \mathbf{c}_t
$\tilde{\mathbf{c}}_t \in \mathbb{R}^{ \mathbf{c}_t }$	Weighted convolutional embedding
$\mathbf{v}_t \in \mathbb{R}^{N_r + K * N_c}$	Visit embedding at t -th timestep
$\mathbf{h}_t \in \mathbb{R}^{N_h}$	Hidden state of GRU at t -th timestep

Problem Formulation

We assume that a patient overall visits clinic T times, generating time-ordered EMR records denoted as $\mathbf{r}_t \in \mathbb{R}^{N_r}$ ($t = 1, 2, \dots, T$). Each EMR record contains N_r features such as different lab test results. Thus the prediction problem in this paper can be formulated as, given t historic EMR data of a patient, i.e., $(\mathbf{r}_1, \dots, \mathbf{r}_t)$, how to predict the patient’s healthcare status y_t which is the probability of suffering from the specific risk (e.g., mortality risk, disease diagnosis, decompensation). The next section will detail our solution AdaCare.

Methods

Figure 4 shows the model structure of AdaCare, a Gated Recurrent Units (GRU) based architecture is used to embed the health status at each clinical visit and perform the healthcare prediction. The visit record sequence is embedded by the GRU to obtain the hidden state \mathbf{h}_t . On the one hand, if the model depends on the latest record \mathbf{r}_t alone, it would be overly sensitive to abnormal values of \mathbf{r}_t brought by the missing data and noise of EMR, thus the prediction may lack robustness. On the other, if the model only depends on the historical characteristics, the alertness of its prediction will be compromised. In AdaCare, both of the historical characteristics \mathbf{c}_t extracted by multi-scale dilated

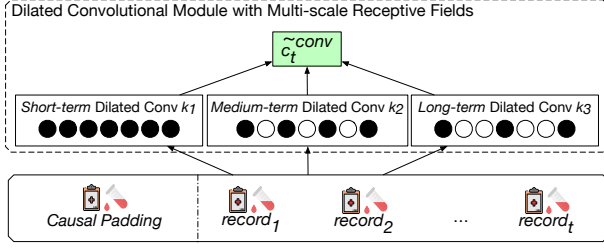


Figure 3: Capture the variation pattern of biomarkers in diverse time scales by dilated convolutional module with multi-scale receptive fields.

convolutional module and the r_t are taken into consideration to build the visit embedding v_t . Finally, we use h_t to predict \hat{y}_t . In summary, the novelty of AdaCare lies in the following two model components:

1. As Figure 3 shows, we develop a dilated convolution (Yu and Koltun 2015) with multi-scale receptive fields to capture the variation characteristics of biomarkers (addressing I_1).
2. As illustrated in Figure 4, we extend the squeeze and excitation block (Hu, Shen, and Sun 2018) to dynamically capture the clinical features which strongly indicate the health risk (addressing I_2).

Multi-Scale Dilated Convolution

One of our goals is to capture the dynamic variations of biomarkers over time and extract such local patterns as additional clinical features. But RNN module alone is difficult to achieve this. The work (Yu and Koltun 2015) demonstrated the effectiveness of *Dilated Convolution* to extract local patterns on images. Thus, AdaCare adopts a similar idea by adding a convolution filter before GRUs. But different to (Yu and Koltun 2015), we extend the dilated convolutional layers with different time spans (i.e., receptive fields), as depicted in Figure 4. By doing so, AdaCare demonstrates the remarkable capability to capture both long-term trends and short-term abnormal variations of biomarkers simultaneously. Mathematically, the *Dilated Convolution* is a convolution applied to input with defined gaps:

$$d[i] = \sum_{l=1}^L b[i + k \cdot l] \cdot z[l], \quad (1)$$

where b is the input biomarkers of records, $d[i]$ is the output feature map, $z[l]$ denotes the convolutional filter of length L , and k corresponds to the dilation rate. We use multiple filters to generate different filter maps, and the number of filter maps is N_c . We concatenate the multiple filter maps to get the final convolution output, denoted as $conv_t^k$. Figure 5 shows how dilated convolutional Layers with different dilation rates work.

In AdaCare, the dilated convolutional module is extended with multi-scale receptive fields consists of multiple parallel convolutional branches with the same filter size and stride but different dilation rates of k_1, k_2, \dots, K . For a given

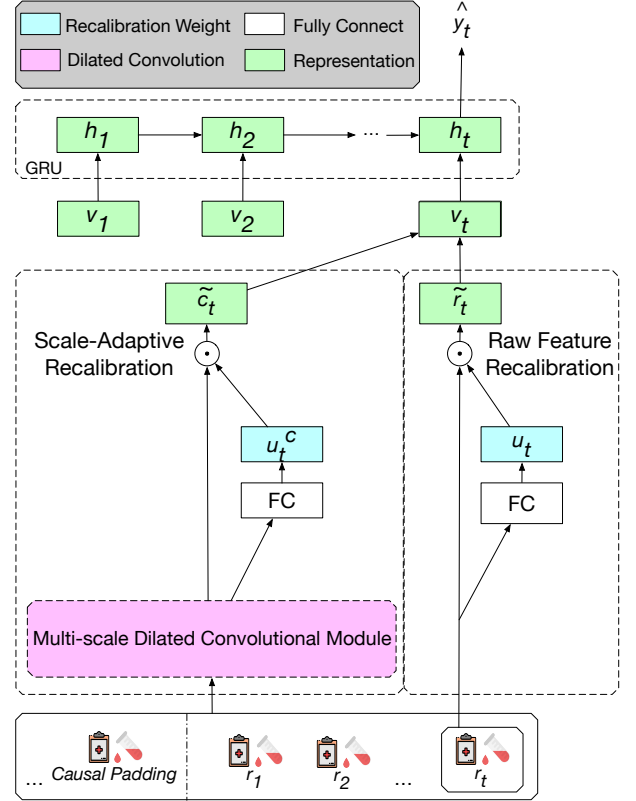


Figure 4: AdaCare Framework. Historical variation patterns of the biomarker in diverse time scales are extracted by a scale-adaptive dilated convolutional module. Both historical characteristics and the most recent record, which is attended by feature recalibration, are combined as clinical visit health status embedding. AdaCare can adaptively capture the predictive feature extracted in proper time scales and indicate the high-risk for the current health condition.

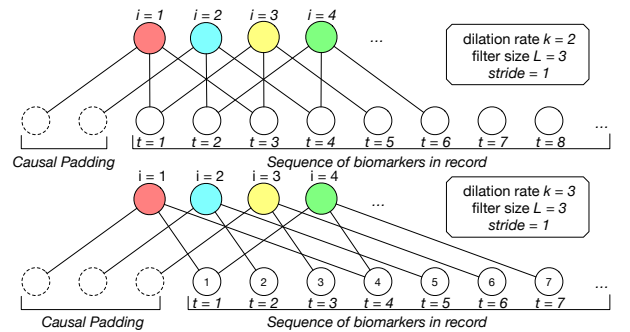


Figure 5: Dilated Convolutional Layers: given the EMR sequence, dilation rate $k = 1$ is a short-term standard convolution, and $k = 2$ means medium-term convolution skipping one record per input, and $k = 3$ means relative long-term convolution skipping two records

dilation rate, each layer takes multiple records of a time span, and the filters scan across the records to generate feature maps, which are concatenated to represent the long and short-term variation by $\mathbf{c}_t = [\text{conv}_t^{k_1}; \text{conv}_t^{k_2}; \dots; \text{conv}_t^K]$. Moreover, to prevent the leakage of follow-up records, we utilize causal padding (Yu and Koltun 2015).

Scale-Adaptive Clinical Feature Recalibration

As we extract the sophisticated variation patterns of biomarkers, the multi-scale dilated convolutional module also introduces redundancy into the model inevitably. Besides, some of the clinical features recorded in EMR have a high correlation with each other or even contribute little to the prediction target. It will reduce the interpretability and robustness of the learned representation if such redundant information is fed into the network. To improve the adaptability of AdaCare in terms of feature utilization, we design the scale-adaptive clinical feature recalibration module based on *SEblock* (Hu, Shen, and Sun 2018). This module is trained to model the nonlinear dependencies between clinical features explicitly. For a particular patient, it can selectively give more weights to the representative and predictive features but suppresses the unimportant features. For patients with different disease conditions (e.g., suffering from chronic disease), the representative and predictive features in the corresponding time scale (e.g., the long-term dilated convolutional feature) would be enhanced. Concretely, we design two fully-connected layers to learn the abstract weight representation and then re-scale it to the original dimension.

$$\mathbf{u}_t^c = \sigma(\mathbf{U}_c \text{ReLU}(\mathbf{W}_c \mathbf{c}_t)), \quad (2)$$

where \mathbf{c}_t is the input of the abstraction operation; parameter matrix $\mathbf{W}_c \in \mathbb{R}^{\lfloor \frac{|c_t|}{r} \rfloor \times |c_t|}$, $\mathbf{U}_c \in \mathbb{R}^{|c_t| \times \lfloor \frac{|c_t|}{r} \rfloor}$; r is the compress ratio that determines the abstraction degree of features; \mathbf{U}_c denotes the mapping matrix which rescales the input into N_c -dimensional; σ is the *Sigmoid* activation function. Then the learned weight \mathbf{u}_t^c can be applied to the original features with an element-wise multiplication:

$$\tilde{\mathbf{c}}_t = \mathbf{u}_t^c \odot \mathbf{c}_t \quad (3)$$

The original input vector \mathbf{c}_t is filtered to be sparser, and the redundancy of the network is reduced. Such feature recalibration can be adjusted adaptively and dynamically through the visits according to the particular health condition.

Besides, in AdaCare, both the raw features and the features captured by the dilated convolutional layers are used to represent the current health status together via a recalibrated skip-connection. The selectively enhanced predictive features can be treated as a precursor of health risk for the given patient.

$$\mathbf{u}_t^r = \sigma(\mathbf{U}_r \cdot \text{ReLU}(\mathbf{W}_r \cdot \mathbf{r}_t)), \quad (4)$$

$$\tilde{\mathbf{r}}_t = \mathbf{u}_t^r \odot \mathbf{r}_t. \quad (5)$$

The weighted raw features and the weighted convolutional features are concatenated together: $\mathbf{v}_t = [\tilde{\mathbf{r}}_t; \tilde{\mathbf{c}}_t]$. The visit embedding \mathbf{v}_t are fed into GRU to obtain hidden representations: $\mathbf{h}_t = \text{GRU}(\mathbf{h}_{t-1}, \mathbf{v}_t)$. The attention mechanism can be easily adopted on the hidden representation here, but it is

not the primary concern in this paper. Finally, the healthcare prediction is obtained as:

$$\hat{y}_t = \sigma(\mathbf{W}_y \mathbf{h}_t + \mathbf{b}) \quad (6)$$

Experiment

We conduct the decompensation prediction experiment on the MIMIC-III dataset and mortality prediction experiment on the ESRD (i.e., end-stage renal disease) dataset. The effectiveness of dynamic feature recalibration is investigated by an overall observation. In order to intuitively show the implication and prediction process of AdaCare, we also develop a simple visualization prototype. The source code of AdaCare, statistics of datasets, case studies, and the visualization prototype are available at the GitHub repository⁴.

Data Preprocessing and Prediction Tasks

- MIMIC-III Dataset**⁵. We use ICU data from the publicly available Medical Information Mart for Intensive Care (MIMIC-III) database (Johnson et al. 2016). We perform the detection of physiologically decompensating patients, which is formulated as a binary classification problem based on patients' clinical events produced during ICU stays (Harutyunyan et al. 2017). Physiologic decompensation is formulated as a problem of predicting if a patient would die within the next 24 hours by continuously monitoring the patient within fixed time-windows. Decompensation labels were curated based on the occurrence of the patient's date of death (DOD) within the next 24 hours, and only about 4.2% of samples are positive in the benchmark. There are 1,203 patients (about 2.89%) with overlong sequences (i.e., > 400). Without loss of fairness, we truncate the length of samples to a reasonable limit (i.e., 400). Eventually, a cohort of 41,602 unique patients with a total of 3,431,622 samples (i.e., records) is used in our dataset. We fix a test set of 15% of patients and divide the rest of the dataset into the training set and validation set with a proportion of 85%:15%. We resample the test set 1000 times using the bootstrap method (Harutyunyan et al. 2017) and calculate the standard deviation of the results.

- ESRD Dataset**. We perform the mortality risk prediction on a real-world end-stage renal disease dataset. In this study, all end-stage renal disease patients who received therapy from January 1, 2006, to March 1, 2018, in a real-world hospital are included to form this dataset. We select the features that are observed in more than 60% patients' records. For missing values, we fill the missing front cells with the data backward to prevent the leakage of future information. If the backward record of a patient is missing, we impute it with the first front observed record of the patient. The cleaned dataset consists of 656 patients with static baseline information and 13,091 dynamic records. There are 1196 records with positive labels

⁴<https://github.com/Accountable-Machine-Intelligence/AdaCare>

⁵<https://mimic.physionet.org>

(i.e., died within 12 months) and 10,804 records with negative labels. We evaluate the models with a 10-fold cross-validation strategy and report the average performance, similar to (Ma, Xiao, and Wang 2018).

Similar to related researches, we assess performance using area under the precision-recall curve (AUPRC) (Keilwagen, Grosse, and Grau 2014), the minimum of precision and sensitivity $\text{Min}(\text{Se}, \text{P}+)$, and area under the receiver operating characteristic curve (AUROC) (Hanley and McNeil 1982). The $\text{Min}(\text{Se}, \text{P}+)$ is calculated as the maximum of $\min(\text{sensitivity}, \text{precision})$ on the precision-recall curve.

Implementation Details and Baselines

Several models share part of the similar insights with AdaCare to learn the representation of patient status, some of which are taken as baseline approaches and listed as follows. We conduct a grid search over hyper-parameters space for the models.

- **GRU** is the standard Gated Recurrent Unit network.
- **RETAIN** (Choi et al. 2016) utilizes a two-level neural attention mechanism to detect influential visits and significant variables, which provide interpretability.
- **T-LSTM** (Baytas et al. 2017) handles irregular time intervals by enabling time decay. We modify it into a supervised learning model.
- **SAnD*** (Song et al. 2018) models clinical time-series data solely based on self-attention. We re-implement SAnD by using $r_{t-k+1:t}$ to build input embedding at the measurement position t (i.e., causal padding (Van Den Oord et al. 2016)), instead of the one proposed in the original paper $r_{t:t+k-1}$, to avoid the violation of causality. The kernel size of convolutional embedding is set to 1.

We also compare AdaCare with the variants of our approaches. Subscript c in Table 2 denotes the multi-scale dilated convolution. Subscripts σ and ς denote the raw feature recalibration module learned with activation function *sigmoid* and *sparsemax* respectively.

For AdaCare, we set the hidden units of AdaCare to 64 for the ESRD dataset and 128 for MIMIC-III dataset. We use 64 filters for convolutional layers, the kernel size is set to 2, and the dilation rate is set to 1,2,3/1,3,5 for ESRD/MIMIC dataset, respectively. For the feature recalibration block, we set the compression ratio to 2/4 for ESRD/MIMIC dataset, respectively. We also utilize the dropout strategy (the dropout rate is 0.5) between the RNN layer and the final output layer for all approaches. We utilize Adam optimizer (Kingma and Ba 2014) with the mini-batch of 128 patients, and the learning rate is set to $1e-3$. The training is done on a machine equipped with CPU: Intel Xeon E5-2630, 256GB RAM, and GPU: Nvidia Titan V. We implement AdaCare with Pytorch 1.1.0.

Results of Healthcare Prediction

Table 2 shows that the performance of all approaches on two datasets: MIMIC-III and the ESRD dataset. AdaCare outperforms all baseline models across both datasets in all eval-

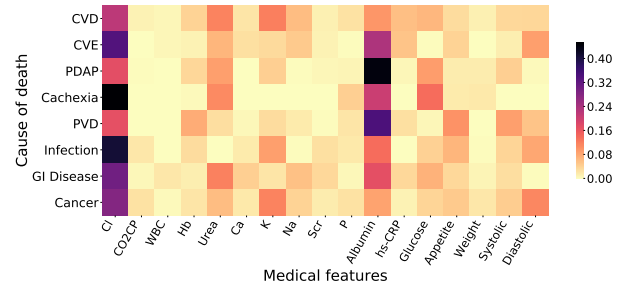


Figure 6: Importance of biomarkers differentiated by dynamic feature recalibration.

uation metrics, especially AUPRC, which is the most informative and the primary evaluation metric when dealing with a highly imbalanced and skewed dataset (Davis and Goadrich 2006; Choi et al. 2018) like the real-world EMR data. Compared to the best baseline model, AdaCare achieves relative improvements of 14.2% and 9.1% for AUPRC in ESRD and MIMIC dataset respectively. Although RETAIN can provide interpretability, its performance is worse than the basic GRU model on both two datasets, which is consistent with the results reported in (Ma et al. 2018a).

AdaCare_{c,-} (i.e., with multi-scale dilated convolution) outperforms the baselines mentioned above. It confirms our assumption that extracting the historical variation pattern of the biomarker in different time scales can depict the health status more comprehensively. AdaCare_{c,σ} (i.e., multi-scale dilated convolution and feature recalibration with *sigmoid* activation function) outperforms the baseline approaches including AdaCare_{c,-}. It suggests that the feature recalibration module can enhance the predictive feature to build the representation effectively, and improve the performance.

To further verify the effectiveness of the model when clearly providing the most high-risk clinical feature in latest visit for the physician, we also test the performance of AdaCare_{c,ς}, which utilizes *sparsemax* as the activation function of the raw feature recalibration. Such recalibration enhances only a few most predictive features and suppresses most of the features. When performing the mortality prediction on the ESRD dataset, the performance of AdaCare_{c,ς} is slightly worse than AdaCare_{c,σ}, and consistently better than most of the comparative approaches. The result indicates that the AdaCare_{c,ς} is still reliable when performing health prediction on the ESRD dataset.

Interpretability and Implications

The case study of a specific sample is usually used to verify the interpretability in the EMR analysis researches, but it is still not convincing enough due to the contingency of case studies. In order to quantitatively identify the reasonability of feature recalibration from an overall perspective, we calculate the average importance weights of biomarkers on different causes of death on ESRD validation sets. The feature-death reason importance are visualized on Figure 6. Some of the essential medical knowledge learned by AdaCare are summarized as follows:

Table 2: Results of Health Risk Prediction. Values in the parentheses denote standard deviations.

	Mortality Prediction on ESRD			Decompensation Prediction of MIMIC		
	AUPRC	min(Se, P+)	AUROC	AUPRC	min(Se, P+)	AUROC
GRU	27.14% (.025)	31.66% (.030)	80.66% (.013)	27.84% (.003)	32.60% (.004)	89.83% (.003)
RETAIN	26.18% (.021)	29.98% (.033)	79.25% (.027)	25.97% (.004)	29.00% (.005)	87.64% (.002)
T-LSTM	27.84% (.019)	33.37% (.028)	81.13% (.021)	26.11% (.003)	31.86% (.004)	89.44% (.002)
SAnD*	26.31% (.033)	29.94% (.037)	79.54% (.032)	25.24% (.003)	28.99% (.004)	88.25% (.003)
AdaCare _{-,σ}	27.66% (.025)	30.55% (.039)	79.64% (.028)	28.11% (.002)	32.71% (.003)	89.77% (.003)
AdaCare _{c,-}	30.77% (.021)	33.45% (.022)	81.22% (.012)	28.37% (.003)	33.10% (.003)	89.81% (.002)
AdaCare _{c,s}	30.98% (.025)	33.31% (.033)	80.61% (.019)	28.95% (.004)	34.23% (.004)	89.93% (.002)
AdaCare _{c,σ}	31.79% (.020)	34.46% (.030)	81.51% (.017)	30.37% (.004)	34.29% (.004)	90.04% (.003)

- Serum albumin is strongly related to adverse outcomes of ESRD patients, especially for the Peritoneal Dialysis-Associated Peritonitis (PDAP). This is consistent with the medical researches (Blake et al. 1993; Spiegel et al. 1993; Cheng et al. 2008; Meijers et al. 2008), which figure out that the abnormal value and decreasing trend of serum albumin usually indicate that the patient may suffer from inflammation and fluid overload.
- Urea is related to gastrointestinal (GI) disease and cachexia. According to medical research (Honda et al. 2006), the abnormal value and decreasing trend of Urea usually indicate that the patient may suffer from low protein intake and malnutrition.
- Serum chlorine (Cl) is strongly related to adverse outcomes of ESRD patients, especially for the cachexia and infection. According to medical experience, the serum chlorine level reflects the renal function of the patient to some extent. However, the relationship between infection, cachexia, and serum chlorine has not been fully explored by existing medical researches. This noteworthy medical finding has already raised the interest of medical experts.

We conduct application-grounded evaluation (Doshi-Velez and Kim 2017) by inviting 12 experienced medical practitioners (with 5-15 years practicing time) from nephrology departments of 5 different hospitals, to evaluate the agreement degree of interpretability generated by AdaCare. The interpretability provided by AdaCare is highly consistent with the practice experience of human experts. Some of the extracted medical knowledge has already been introduced as the ESRD management aid by physicians. More details about the experiment are described at our GitHub repository.

Conclusion

In this paper, AdaCare is proposed to learn the clinical health status representation. Specifically, we utilize the dilated convolutions with multi-scale receptive fields to capture the long and short-term historical variation of biomarkers. AdaCare models the nonlinear dependencies of features by extending SE-block. Such a feature re-calibration process selectively enhances predictive features extracted in proper time scales and the most high-risk factors in the latest visit. It builds effective health status representation and provide reasonable interpretability. Experiment results on

MIMIC-III dataset and ESRD dataset show that AdaCare outperforms the baseline approaches with powerful interpretability. Medical knowledge learned by AdaCare has been positively confirmed by human medical experts and related medical literature.

Acknowledgments

This work is supported by National Science and Technology Major Project (No. 2018ZX10201002) and the fund of Peking University Health Science Center (BMU20160584). WR is supported by ORCA PRF Project (EP/R026173/1).

References

- Anna Malkina. 2018. Chronic kidney disease. [Online; October 2018].
- Bai, T.; Zhang, S.; Egleston, B. L.; and Vucetic, S. 2018. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 43–51. ACM.
- Baytas, I. M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A. K.; and Zhou, J. 2017. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 65–74. ACM.
- Bharadwaj, S.; Ginoya, S.; Tandon, P.; Gohel, T. D.; Guirguis, J.; Vallabh, H.; Jevonn, A.; and Hanouneh, I. 2016. Malnutrition: laboratory markers vs nutritional assessment. *Gastroenterology report* 4(4):272–280.
- Blake, P. G.; Flowerdew, G.; Blake, R. M.; and Oreopoulos, D. G. 1993. Serum albumin in patients on continuous ambulatory peritoneal dialysis—predictors and correlations with outcomes. *Journal of the American Society of Nephrology* 3(8):1501–1507.
- Cheng, L.-T.; Tang, L.-J.; Chen, H.-M.; Tang, W.; and Wang, T. 2008. Relationship between serum albumin and pulse wave velocity in patients on continuous ambulatory peritoneal dialysis. *Vascular health and risk management* 4(4):871.
- Cheng, Y.; Wang, F.; Zhang, P.; and Hu, J. 2016. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, 432–440. SIAM.

- Choi, E.; Bahadori, M. T.; Sun, J.; Kulas, J.; Schuetz, A.; and Stewart, W. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, 3504–3512.
- Choi, E.; Xiao, C.; Stewart, W.; and Sun, J. 2018. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *Advances in Neural Information Processing Systems*, 4547–4557.
- Davis, J., and Goadrich, M. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240. ACM.
- Doshi-Velez, F., and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Hanley, J. A., and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143(1):29–36.
- Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; and Galstyan, A. 2017. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*.
- Honda, H.; Qureshi, A. R.; Heimbürger, O.; Barany, P.; Wang, K.; Pecoits-Filho, R.; Stenvinkel, P.; and Lindholm, B. 2006. Serum albumin, c-reactive protein, interleukin 6, and fetuin a as predictors of malnutrition, cardiovascular disease, and mortality in patients with esrd. *American Journal of Kidney Diseases* 47(1):139–148.
- Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.
- Isakova, T.; Xie, H.; Yang, W.; Xie, D.; Anderson, A. H.; Scialla, J.; Wahl, P.; Gutiérrez, O. M.; Steigerwalt, S.; He, J.; et al. 2011. Fibroblast growth factor 23 and risks of mortality and end-stage renal disease in patients with chronic kidney disease. *Jama* 305(23):2432–2439.
- Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3:160035.
- Keilwagen, J.; Grosse, I.; and Grau, J. 2014. Area under precision-recall curves for weighted and unweighted data. *PLoS one* 9(3):e92209.
- Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Lee, W.; Park, S.; Joo, W.; and Moon, I.-C. 2018. Diagnosis prediction via medical context attention networks using deep generative modeling. In *2018 IEEE International Conference on Data Mining (ICDM)*, 1104–1109. IEEE.
- Ma, F.; Chitta, R.; Zhou, J.; You, Q.; Sun, T.; and Gao, J. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1903–1911. ACM.
- Ma, F.; Gao, J.; Suo, Q.; You, Q.; Zhou, J.; and Zhang, A. 2018a. Risk prediction on electronic health records with prior medical knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1910–1919. ACM.
- Ma, Y.; Wang, Y.; Wang, H. Y.; Li, Y.; Zhou, Z. J.; Liu, B. Y.; Li, X. M.; and Chen, L. M. 2018b. Predictors of short-term outcomes of patients with peritoneal dialysis-associated peritonitis. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao Acta Academiae Medicinae Sinicae* 40(1):13.
- Ma, L.; Zhang, C.; Wang, Y.; Ruan, W.; Wang, J.; Tang, W.; Ma, X.; Gao, X.; and Gao, J. 2020. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.
- Ma, T.; Xiao, C.; and Wang, F. 2018. Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, 261–269. SIAM.
- Meijers, B. K.; Bammens, B.; Verbeke, K.; and Evenepoel, P. 2008. A review of albumin binding in ckd. *American journal of kidney diseases* 51(5):839–850.
- Rabkin, S. W.; Mathewson, F. A.; and Tate, R. B. 1978. Predicting risk of ischemic heart disease and cerebrovascular disease from systolic and diastolic blood pressures. *Annals of internal medicine* 88(3):342–345.
- Song, H.; Rajan, D.; Thiagarajan, J. J.; and Spanias, A. 2018. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Spiegel, D. M.; Anderson, M.; Campbell, U.; Hall, K.; Kelly, G.; McClure, E.; and Breyer, J. A. 1993. Serum albumin: a marker for morbidity in peritoneal dialysis patients. *American journal of kidney diseases* 21(1):26–30.
- Sun, M.; Tang, F.; Yi, J.; Wang, F.; and Zhou, J. 2018. Identify susceptible locations in medical records via adversarial attacks on deep predictive models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 793–801. ACM.
- Tangri, N.; Ansell, D.; and Naimark, D. 2011. Determining factors that predict technique survival on peritoneal dialysis: application of regression and artificial neural network methods. *Nephron Clinical Practice* 118(2):c93–c100.
- Valko, M., and Hauskrecht, M. 2010. Feature importance analysis for patient management decisions. *Studies in health technology and informatics* 160(Pt 2):861.
- Van Den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. *CoRR abs/1609.03499*.
- Yu, F., and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.