GRASP: Generic Framework for Health Status Representation Learning Based on Incorporating Knowledge from Similar Patients

Chaohe Zhang[1,3], Xin Gao[1,3], Liantao Ma[1,3], Yasha Wang[1,2*], Jiangtao Wang[4] and Wen Tang[5]

[1]Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing, China
[2]National Engineering Research Center of Software Engineering, Peking University, Beijing, China
[3]School of Electronics Engineering and Computer Science, Peking University, Beijing, China
[4]The Centre for Intelligent Healthcare, Coventry University, UK
[5]Division of Nephrology, Peking University Third Hospital, Beijing, China
{wangyasha, choc}@pku.edu.cn, {jiangtao.wang}@coventry.ac.uk

Abstract

Deep learning models have been applied to many healthcare tasks based on electronic medical records (EMR) data and shown substantial performance. Existing methods commonly embed the records of a single patient into a representation for medical tasks. Such methods learn inadequate representations and lead to inferior performance, especially when the patient's data is sparse or low-quality. Aiming at the above problem, we propose GRASP, a generic framework for healthcare models. For a given patient, GRASP first finds patients in the dataset who have similar conditions and similar results (i.e., the similar patients), and then enhances the representation learning and prognosis of the given patient by leveraging knowledge extracted from these similar patients. GRASP defines similarities with different meanings between patients for different clinical tasks, and finds similar patients with useful information accordingly, and then learns cohort representation to extract valuable knowledge contained in the similar patients. The cohort information is fused with the current patient's representation to conduct final clinical tasks. Experimental evaluations on two real-world datasets show that GRASP can be seamlessly integrated into state-of-the-art models with consistent performance improvements. Besides, under the guidance of medical experts, we verified the findings extracted by GRASP, and the findings are consistent with the existing medical knowledge, indicating that GRASP can generate useful insights for relevant predictions.

Introduction

With the rapid growth and accumulation of electronic medical records (EMR) data, deep learning methods have been widely applied in many healthcare tasks, such as mortality prediction, patients subtyping, and diagnosis prediction. These methods can assist doctors in analyzing patients' health status, formulating reasonable treatment, and preventing adverse outcomes in a more intelligent and effective way.

EMR data are temporally sequenced by patient clinical records that are represented by a set of medical variables. Most existing methods embed the EMR data of each single patient into a representation separately and perform medical tasks based on it (Baytas et al. 2017; Choi et al. 2018; Ma et al. 2020b). However, EMR data are usually sparse (Xu et al. 2018), and while dealing with a patient record with low quality, such methods will learn inadequate representations and lead to inferior performance. Thus, some researchers try to enhance the performance by incorporating external information. For example, GRAM (Choi et al. 2017) and KAME (Ma et al. 2018b) incorporate the ontologies of the medical codes. Their essence is to incorporate external information beyond the dataset to learn a better representation for the patient, and they achieve improvements in some conditions. However, these approaches do not work well when people hardly obtain external information or prior knowledge about them, especially for some rare diseases or emerging diseases (e.g., COVID-19) (Huang et al. 2020). Furthermore, such ontology information is often not applicable due to the idiosyncratic use of terminology (Choi et al. 2018). A challenge arises now, that is, how to fully utilize such EMR data to learn adequate patient representations without external knowledge?

In fact, in addition to the methods of using external information, fully mining the correlation between similar patients can also improve the performance. This intuition is based on the observation of how human doctors use the similarity between patients to assist the clinical analysis. When a patient

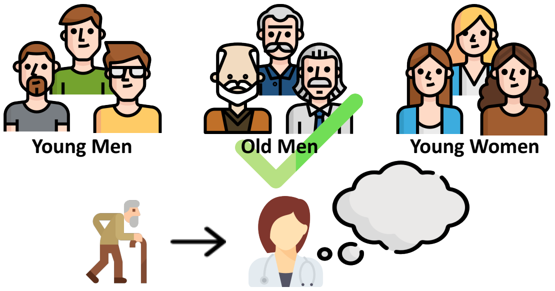
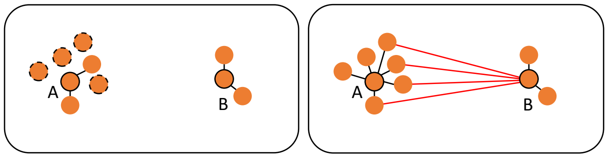


Figure 1: The similar patients can provide auxiliary information for current analysis and treatment.

[*]Corresponding Author



(a) K = 2

(b) K = 6

Figure 2: Two Dilemmas of Selecting Similar Patients

goes to see a doctor, the doctor will first examine the lab test results. Then, as shown in Figure 1, the doctor usually recalls the health status of the similar patients that she/he has treated, or looks up their records from the hospital system, and then assess and treat the current patient. The same insight can be used in deep learning models for healthcare tasks. While processing the current patients, there are other cases with conditions alike. The information of these similar patients can be utilized as guidance for the current prognosis.

Although seeming straightforward, applying this intuition to real clinical tasks will face the following challenges:

Challenge 1. How to measure the similarity between patients? (Zhu et al. 2016) proposes a similarity evaluation model based on the temporal matching of patient EMR. However, they did not associate the connotation of similarity with clinical tasks. In different clinical tasks (e.g., mortality prediction and different disease diagnosis), the patient characteristics that need attention are different, so two patients who are considered similar in one clinic task may not be considered so similar in another. Furthermore, different tasks correspond to different healthcare models. Therefore, how to design a unified framework that can consider differences in clinical tasks with different models and reasonably measure the similarity between patients is the first challenge.

Challenge 2. How to select similar patients? (Suo et al. 2018) uses the intuitive K nearest neighbor method, that is, for any given patient A, finding the nearest (most similar) K patients as the similar patients. However, this idea does not work well in a space with uneven data distribution. As shown in Figure 2, there are more samples similar to A, but less similar to B. When K is small (Figure 2-a), there are many similar patients around A that are not fully utilized (dotted circles). However, when increasing K (Figure 2-b), some patients who are not similar to B are also selected as the similar patients (red lines), resulting in a negative effect.

Challenge 3. How to incorporate the auxiliary information from the similar patients? The amounts of auxiliary information required by various types of patients are different. Some patients have sufficient data, and their health status representations are relatively easy to extract. Thus they need less knowledge from similar patients and rely more on their own. On the contrary, for other patients whose representations are hard to extract, the more auxiliary information is needed. Therefore, it is worth thinking about how to adaptively fuse the current patient information with the auxiliary information.

By jointly considering the above issues, we propose a generic framework, GRASP, which can be integrated with existing healthcare models. Our main contributions are summarized as follows:

- We propose a generic framework called GRASP, which boosts the performances of existing healthcare models by fully considering both the current patient's information and the auxiliary information from similar ones. (Response to Challenge 1)
- Specifically, GRASP automatically assigns different types of similar patients into cohorts and extracts cohort representations. Considering the interdependency of the cohorts, the representations are formed as a graph, and GNN is used to extract the enhanced cohort representation as the auxiliary information. (Response to Challenge 2)
- Next, GRASP assigns the weights of the auxiliary representation and the patient representation, and adaptively fuses them to depict the patient more comprehensively. (Response to Challenge 3) Besides, in this way, the learned patient representations are facilitated to be discriminative group-wisely.
- Extensive experiments show that our framework can be seamlessly integrated into state-of-the-art models with a consistent performance improvement under various settings. Besides, the findings discovered by GRASP are in accord with experts and medical knowledge, which shows it can provide useful insights and explanations.

Related Work

Over the past years, deep learning models have shown the capability to perform mortality prediction (Suresh, Gong, and Guttag 2018; Tan et al. 2020; Ma et al. 2020a,b), patients subtyping (Baytas et al. 2017), and diagnosis prediction (Lee et al. 2018; Ma, Xiao, and Wang 2018; Ma et al. 2017; Gao et al. 2019). Though the medical tasks vary from each other, their essences are extracting the health status representations of patients. For example, RETAIN (Choi et al. 2016) uses a two-level neural attention model to detect influential visits and significant variables. T-LSTM (Baytas et al. 2017) handles irregular time intervals by enabling time decay to learn better patient representations. Concare (Ma et al. 2020b) embeds the feature sequences separately and uses the self-attention to capture the healthcare context to learn personalized representations. Furthermore, some researches incorporate external information to boost the performance. GRAM (Choi et al. 2017) and KAME (Ma et al. 2018b) incorporate medical ontologies to train the model sufficiently. The fundamental idea of them is to aggregate external information to enhance representation learning for the final tasks.

In fact, in addition to the method of using external information, fully extracting the correlation between similar patients inside the dataset can also improve the performance of the model. There are some patients with similar status, and they are more likely to suffer from a similar outcome. To this end, some researchers focus on similarity discovery for healthcare. (Zhu et al. 2016) proposes a patient similarity evaluation model based on the temporal matching of patient EMR for cohort study. Moreover, (Suo et al. 2018) collects

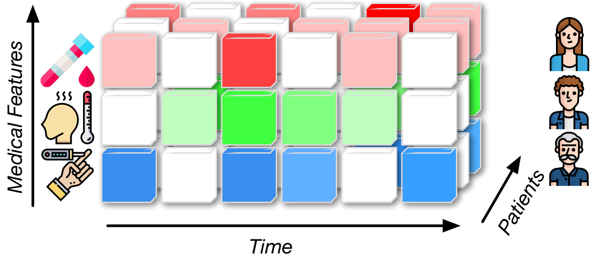


Figure 3: EMR Data Description

and predicts K most similar patients together. They use the most common label appearing as the current patient's predicted label. However, as the discussion in Challenge 1 and 2, they do not work very well in some conditions. Different from the above models, we propose GRASP, which can explicitly capture the holistic observation among similar patients and incorporate such auxiliary information to learn a more comprehensive representation.

Preliminary

Electronic Medical Records (EMR) data record the medical processes of the patients. As shown in Figure 3, a patient has a sequence of records along with time of visits, generating time-ordered EMR records, which are denoted as $\boldsymbol{r}_t \in \mathbb{R}^{N_r}$ $(t = 1, 2, \cdots, T)$. Each EMR record contains N_r medical features (e.g., laboratory measurements).

The predictive problem in this paper can be formulated as given T historic EMR data of a patient, to predict the patient's future health condition y (i.e., prognosis). In general, the future health condition is defined as the probability of suffering from a specific risk (e.g., mortality). Since GRASP is a generic framework, it utilizes existing healthcare models as the $Backbone$ and improves them. We follow the definition to formulate the problem as: $\hat{y} = \texttt{GRASP}(\text{Backbone}(\boldsymbol{r}_1, \cdots, \boldsymbol{r}_T))$.

Methodology

Overview

Figure 4 shows the architecture of GRASP. It comprises the following sub-modules:

- The patient representation extraction module embeds the patient's clinical records into a representation with a backbone model.
- The cohort discovery and utilization module finds the patients with similar health status to the current patient, aggregates them into a cohort, and extracts the guidance representation.
- The adaptive fusion module combines the above-learned representations (i.e., current patient representation and guidance representation) adaptively for the final task.

Patient Representation Extraction

Since GRASP is a generic framework and needs to perform on patient representations, a backbone model is required to work as the representation extractor. Such a backbone extractor can be one of the existing state-of-the-art models (e.g., (Ma et al. 2020b)) and the hidden representations before the final layer of those models are used as the representations of the patients. For ease of understanding, RNN is used to illustrate the process in Figure 4 and here.

Given a sequence of medical records along with visits $\boldsymbol{r}_1, \cdots, \boldsymbol{r}_T$, the representation of the patient can be obtained as: $v_t = \text{ReLU}(W_v r_t)$, $h_t = \text{RNN}(v_t, h_{t-1})$, where $t = 1, ..., T$ is the time steps of the patient's visits. $W_v \in \mathbb{R}^{N_v \times N_r}$ is a weight matrix and we ignore the bias terms for simplicity. $v_t \in \mathbb{R}^{N_v}$ is the obtained visit-level embedding for t-th visit of the patient. $h_t \in \mathbb{R}^{N_h}$ is the hidden state. N_r, N_v, and N_h are the dimensions of records, visit embeddings and hidden states, respectively. We can obtain the final latent hidden state h_T as the representation of the specific patient, which is often used for some prediction tasks. For other backbone models, the process of obtaining the patient representation h_T can be abbreviated as:

$$h_T = \text{Backbone}(\boldsymbol{r}_1, \cdots, \boldsymbol{r}_T). \quad (1)$$

Similar Patient Cohort Discovery and Utilization

Now, the representations are obtained in the previous section. For a specific patient, as discussed before, the knowledge from the similar patients can be utilized as guidance for the analysis or prognosis. A straightforward way to find similar patients is to calculate similarity via the learned representations h_T of every patient pair. However, how to identify the really similar patients from the seemingly similar ones is hard. This will be more challenging for those unbalanced datasets. For instance, as shown in Figure 2, there are more samples similar to A, but less similar to B. If we select k *similar* patients for the current patient and k is set small, as shown in Figure 2-a, many samples similar to A cannot be selected. Thus, the information about the similar patients cannot be fully utilized. If k is set large, as shown in Figure 2-b, some samples that are not similar to B will also be selected, resulting in a negative effect. Thus, we argue that collecting each kind of similar patients into cohorts (i.e., assigning the number of similar samples automatically) is a more robust way.

For every batch of samples, the patients' representations are clustered via K-Means (Jain 2010) with Euclidean distance. Then, the centroids of each cluster are extracted to form a centroid (i.e., prototype) matrix $\Gamma \in \mathbb{R}^{N_c \times N_h}$, where N_c is the number of cohorts. K-Means does not change the points in the feature space, so the process is differentiable.

Next, in general, a direct way to select which cluster the current patient belongs to is referring to the result of K-Means. However, in the early training phase, when the model is not convergent and the representations are not fully learned, the result is unstable. Thus, considering the exploration-exploitation decisions[1] in reinforcement learning, we need exploration to find better cluster attribution in the earlier training phase and exploitation to maintain the

[1] Exploration, where we gather more information that might lead to better decisions in the future. Exploitation, where we make the best decision given current information.

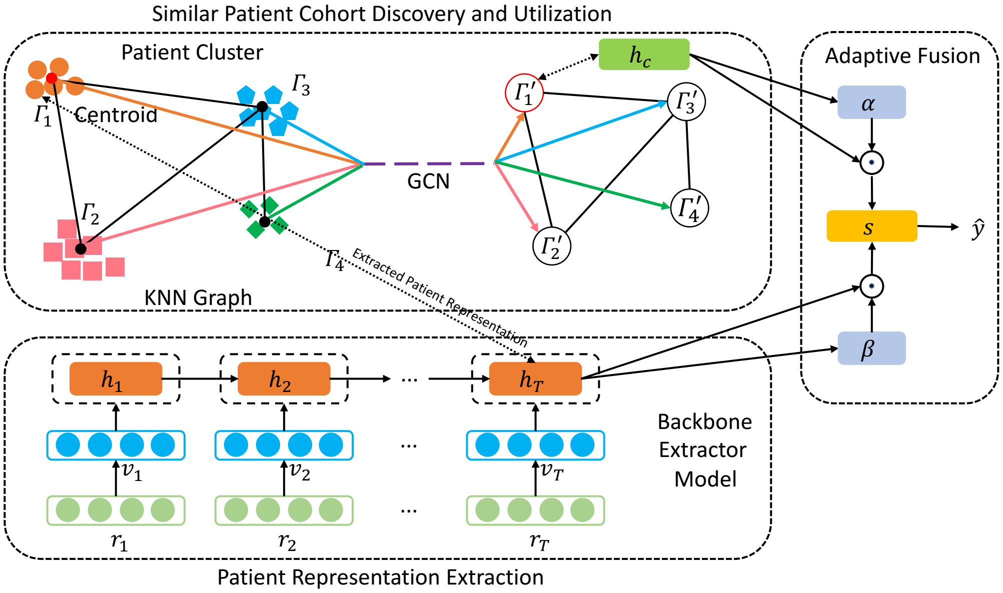


Figure 4: The GRASP framework

current best decision later (Sutton and Barto 2018). Sampling can solve such a condition (Maddison, Tarlow, and Minka 2014), especially when it is with decay, which can turn exploration to exploitation gradually. So in practice, we can acquire a more robust selection by introducing the Gumbel-Max technique (Gumbel 1954; Maddison, Tarlow, and Minka 2014), which provides an efficient sample approach. The similarity of representation of the current patient and each centroid can be calculated by:

$$e = h_T \Gamma^{\intercal}. \tag{2}$$

Then, the Gumbel noise, treated as a form of regularization, is added to e in Equation 2, and the softmax function is performed. The cluster similarity distribution is calculated:

$$\begin{aligned} g &= -\log(-\log u), \\ \tilde{e} &= (e+g)/\tau, \\ a &= \mathrm{softmax}(\tilde{e}), \end{aligned} \tag{3}$$

where g is the Gumbel noise calculated from a uniform distribution $u \sim \mathcal{U}(0,1)$, and τ is the temperature, which can turn exploration to exploitation gradually during the training process. As τ getting close to 0, the softmax function is similar to the argmax operation, and it becomes uniform distribution gradually when $\tau \to \infty$. When sampling, a hard version is performed to select one cluster per patient, which means the samples will be discretized as one-hot vectors. The above sampling steps can be formulated as:

$$a = \text{Gumbel-Softmax}(e), \tag{4}$$

where a is a one-hot vector indicating which cluster the current patient belongs to.

Then, we can use the centroid vector of the cluster as the similar patients' information straightforwardly. Nevertheless, there remain interactions between clusters. For example, if we select a large cohort number, which means dividing patients into fine-grained clusters, some close clusters still share similar characteristics. Such interactions need to be captured. Thus, a K-nearest neighbor (K-NN) graph G is constructed from the centroids of clusters and A is the adjacency matrix of the graph, which shows the connectivity between the K-nearest centroid (cluster) representations. And the self-connection is added to A to make such an aggregation more self-attentive: $\hat{A} = A + I$, where I is the identity matrix and $\hat{A}$ is the adjacency matrix of G with self-connection. Next, graph convolutional layers (GCN) (Kipf and Welling 2016) are applied to enhance the representation learning by leveraging the structural information:

$$\Gamma' = \mathrm{ReLU}\left(\hat{A}\ \mathrm{ReLU}\left(\hat{A}\Gamma W^0\right) W^1\right), \tag{5}$$

where Γ' is the enhanced cluster representations form Γ, and W^0 and $W^1 \in \mathbb{R}^{N_h \times N_h}$ are the projection matrices. The corresponding auxiliary cohort representation is obtained:

$$h_c = a\Gamma'. \tag{6}$$

Adaptive Attention Fusion

Now there are two representations related to the patient: h_T and h_c. The former focuses on the patient herself, while the latter refers to similar patients. An adaptive fusion method is utilized to extract the proper amount of information from them and build a comprehensive patient representation.

More specifically, two weights ($\alpha, \beta \in \mathbb{R}$) are introduced to determine the amount of the above two representations, which are obtained by fully connected layers on h_c and h_T:

$$\alpha = \mathrm{Sigmoid}\left(W_c h_c\right), \tag{7}$$

$$\beta = \mathrm{Sigmoid}\left(W_T h_T\right), \tag{8}$$

where $W_c, W_T \in \mathbb{R}^{1\times N_h}$ are the projection matrices. We add a constraint $\alpha + \beta = 1$ by calculating $\alpha = \frac{\alpha}{\alpha+\beta}$, $\beta = 1 - \alpha$. The final representation can be obtained as:

$$s = \alpha \cdot h_c + \beta \cdot h_T. \tag{9}$$

Then the predictor can be built via a fully connected layer. Mathematically, the predicted probability can be calculated:

$$\hat{y} = \mathrm{Sigmoid}(W_{final} s), \tag{10}$$

where $W_{final} \in \mathbb{R}^{1\times N_h}$ is the weight matrix. The cross-entropy loss is used as the loss function:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}(y_i^{\intercal}\log(\hat{y}_i) + (1-y_i)^{\intercal}\log(1-\hat{y}_i)), \tag{11}$$

where B is the batch size. $\hat{y}_i \in [0,1]$ is the predicted probability, and $y_i \in \{0,1\}$ is the ground truth. In a real-world clinical scene, we can cluster the whole dataset and save the centroids. In this way, when a new patient comes in, the centroids can be used as the substitution of similar patients.

Experiment

Data Description and Task Formulation

Cardiology Dataset We perform the sepsis prediction on the open-source PhysioNet cardiology dataset (Reyna et al. 2019), which were collected from three geographically distinct U.S. hospitals over the past decade. They are labeled by Sepsis-3 clinical criteria. The dataset consists of 40,336 patients and consists of a combination of hourly vital signs and lab values. The dataset is divided into the training set, validation set, and test set with a proportion of 0.8 : 0.1: 0.1. It is an imbalanced dataset and statistics are in Table 1.

CKD Dataset Another dataset we use is a real-world chronic kidney disease (i.e., CKD) dataset. In this study, all CKD patients who received therapy from January 1, 2006, to March 1, 2018, in a real-world hospital are included to form this dataset[2]. The statistics of the CKD dataset are presented in Table 1. The / in Table 1 is used to separate label information of mortality prediction task (left) and disease diagnosis task (right). The latter task will be described in the analysis part. The mortality prediction task on CKD dataset is defined as a binary classification task of predicting the death of a patient in one year. Due to the scarce amount of CKD data, 10-fold cross-validation is employed.

We assess performance using the area under the precision-recall curve (AUPRC), the minimum of precision and sensitivity Min(Se,P+), and F1-score. AUPRC is the most informative evaluation metric when dealing with a highly imbalanced and skewed dataset (Davis and Goadrich 2006; Choi et al. 2018; Chu et al. 2019) like the real-world data.

Experimental Setup and Baselines

To conduct the experiment, we use the Adam optimization with learning rate = 1e-3. More information are available at [3]. To fairly compare different approaches, the hyper-parameters of the models are fine-tuned by grid search on the training data. We include several state-of-the-art models as our baseline models as well as the backbones of GRASP:

Dataset	Cardiology	CKD
# of patients	40,336	662
# of visits	1,552,210	13,108
Avg. # of visits	38.48	19.95
Max. # of visits	336	69
Min. # of visits	8	1
# of features	33	17
% of positive labels	7.26%	38.97% / 36.40%

Table 1: Statistics of the Datasets

- RETAIN (NeurIPS) (Choi et al. 2016) utilizes a two-level attention to detect weights of visits and variables.
- T-LSTM (SIGKDD) (Baytas et al. 2017) handles time intervals by a time decay mechanism in LSTM.
- TimeNet$_*$ (IJCAI) (Gupta et al. 2018) maps clinical time series separately and aggregates all the feature embeddings to conduct healthcare prediction. To conduct a fair comparison, we do not use the pre-trained model.
- ConCare$_*$ (AAAI) (Ma et al. 2020b) embeds the feature sequences separately and uses the self-attention to model dynamic features and static information. For a fair comparison, the static information is not considered here.

We also conduct the following ablation studies:

- GRASP$_{1-}$ does not capture the interaction between clusters. It uses the centroid as the auxiliary information.
- GRASP$_{2-}$ does not have the Gumbel sampling module.

Prediction Results

Table 2 shows the performance of models with GRASP and baselines on the two datasets. For sepsis prediction on the Cardiology dataset, the number in () denotes the standard deviation of bootstrapping for 1000 times. And for mortality prediction on CKD dataset, it denotes the standard deviation of 10-fold cross-validation. We can observe that GRASP consistently increases the performance of all the baselines. Although RETAIN can provide interpretability, the quantitative performance is sacrificed, which is consistent with the results reported in (Ma et al. 2018a). ConCare (Ma et al. 2020b) uses the self-attention to capture the interdependency between features and achieves the best results among the baselines. However, with GRASP, the interdependency among patients is added, which leads to a performance boost. As shown in Table 1, the Cardiology dataset is more imbalanced and sparse than the CKD dataset, and the performance gain of GRASP is larger. This observation also implies that GRASP improves the performance better on such low-quality datasets.

Analysis

Extra information utilization vs. Cause of death To explore how the auxiliary information from other similar patients affects mortality prediction, we further analyze the

[2] This study was approved by the Research Ethical Committee.

[3] https://github.com/choczhang/GRASP

	Sepsis Prediction on Cardiology Dataset			Mortality Prediction on CKD Dataset		
Methods	AUPRC	min(Se, P+)	F1-Score	AUPRC	min(Se, P+)	F1-Score
GRU	0.6771 (0.02)	0.6117 (0.02)	0.6114 (0.01)	0.7126 (0.02)	0.6628 (0.02)	0.6431 (0.01)
GRASP+GRU	**0.7268** (0.02)	**0.6508** (0.01)	**0.6491** (0.01)	**0.7483** (0.01)	**0.6971** (0.01)	**0.6728** (0.02)
RETAIN	0.6580 (0.02)	0.6123 (0.02)	0.6198 (0.01)	0.7063 (0.02)	0.6496 (0.02)	0.6241 (0.02)
GRASP+RETAIN	**0.7003** (0.02)	**0.6485** (0.02)	**0.6394** (0.01)	**0.7256** (0.02)	**0.6715** (0.01)	**0.6604** (0.01)
T-LSTM	0.7138 (0.02)	0.6553 (0.02)	0.6587 (0.02)	0.7180 (0.01)	0.6702 (0.02)	0.6438 (0.01)
GRASP$_{1-}$+T-LSTM	0.7336 (0.02)	0.6758 (0.01)	0.6635 (0.01)	0.7209 (0.01)	0.6851 (0.02)	0.6578 (0.01)
GRASP$_{2-}$+T-LSTM	0.7465 (0.02)	0.6742 (0.01)	0.6702 (0.02)	0.7365 (0.01)	0.6825 (0.02)	0.6636 (0.02)
GRASP+T-LSTM	**0.7513** (0.01)	**0.6821** (0.02)	**0.6798** (0.01)	**0.7496** (0.01)	**0.6986** (0.02)	**0.6751** (0.01)
TimeNet$_*$	0.7570 (0.02)	0.6762 (0.02)	0.6785 (0.01)	0.7358 (0.02)	0.6819 (0.02)	0.6504 (0.01)
GRASP$_{1-}$+TimeNet$_*$	0.7793 (0.02)	0.6791 (0.02)	0.6887 (0.01)	0.7423 (0.02)	0.6998 (0.02)	0.6817 (0.01)
GRASP$_{2-}$+TimeNet$_*$	0.7803 (0.02)	0.6850 (0.02)	0.6849 (0.01)	0.7517 (0.02)	**0.7020** (0.02)	0.6866 (0.01)
GRASP+TimeNet$_*$	**0.7885** (0.02)	**0.6977** (0.02)	**0.7045** (0.01)	**0.7523** (0.02)	0.7013 (0.02)	**0.6885** (0.02)
ConCare$_*$	0.7770 (0.02)	0.7010 (0.02)	0.7056 (0.01)	0.7368 (0.02)	0.6757 (0.02)	0.6600 (0.01)
GRASP$_{1-}$+ConCare$_*$	0.7881 (0.02)	0.7176 (0.02)	0.7185 (0.01)	0.7493 (0.02)	0.6829 (0.02)	0.6814 (0.01)
GRASP$_{2-}$+ConCare$_*$	0.7925 (0.01)	0.7080 (0.02)	0.7169 (0.01)	0.7488 (0.02)	0.6972 (0.02)	0.6902 (0.02)
GRASP+ConCare$_*$	**0.8014** (0.02)	**0.7210** (0.02)	**0.7234** (0.01)	**0.7597** (0.02)	**0.7059** (0.02)	**0.6945** (0.01)

Table 2: Results for the tasks on Cardiology and CKD Datasets

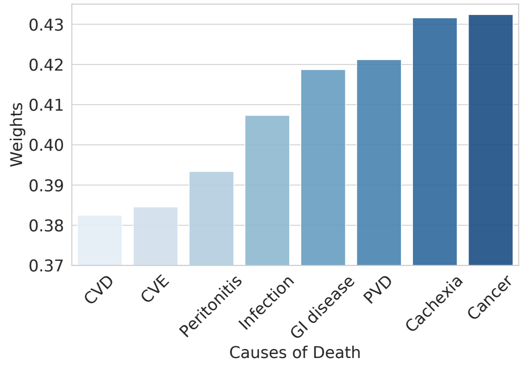


Figure 5: Extra Information Utilization vs. Cause of Death

extra information utilization w.r.t. different causes of death (COD) in the CKD dataset. In the Adaptive Attention Fusion step, GRASP generates two weights. The α determines how much the auxiliary information from other similar patients (i.e., h_c) affects mortality prediction, and the β shows how much effect the information from the patient (i.e., h_T) has. We randomly sample 20% of the patients as the test set and utilize the rest as the training set. After training, the weights of the auxiliary information (i.e., α) are collected on the test set. The average weights of the auxiliary information w.r.t patients with different COD are shown in Figure 5.

The result shows that patients who died of cardiovascular (CVD), cerebrovascular (CVE), and peritonitis have the lowest α , which means that their prediction needs less auxiliary information from similar patients. These diseases are acute diseases (Fried et al. 1996; Kannel et al. 1987). In general, the health status of patients who have these diseases tends to deteriorate rapidly in a short period of time (ESTANOL and M. MARIN 1975), therefore the health status is more exclusive and similar patients are scarce. Thus the information from similar patients has less guidance for prediction, which explains why GRASP generates lower weights.

In contrast, cancer, cachexia, and peripheral vascular diseases (PVD) are relatively chronic diseases. The health status of patients with these diseases change more slowly and more common since their health status often deteriorates more chronically (Prentice and Gloeckler 1978; Derogatis, Abeloff, and Melisaratos 1979). In this way, the guidance from other similar patients is more helpful, since more patients have experienced similar status, which guides the prediction of the current patient. This finding also suggests clinicians pay more attention to patients with CVD and CVE in order to make timely interventions and save more lives.

Patient Cohort Study In this part, following the operation of GRASP, we conduct patient cohort discovery for similar patients on the CKD dataset to investigate the expressive power of the patient representation learned with GRASP. Chronic kidney disease (CKD) is a chronic disease, and the patients need to receive continuous medical analysis for years or even decades. Patient cohort discovery is to seek patient groups with similar disease progression pathways (Baytas et al. 2017) and it can help clinicians develop targeted treatment plans and prevent adverse outcomes.

Cluster Performance. First, we compare the performance of patient clustering for the models with/without GRASP. We use the hidden representations before the final layer of those models as representations for patients' health status. The learned representations are used to cluster the patients by K-Means algorithm (Jain 2010). Since we do not know the ground truth groups, we use Calinski-Harabasz score (Caliński and JA 1974) (C-H score) to evaluate the cluster performance **quantitatively**. A higher C-H score relates to a better method. The C-H score is calculated as: $\text{Calinski-Harabasz score} = \frac{\text{tr}(B_k)}{\text{tr}(W_k)} \frac{m-k}{k-1}$, where m is the sample size, k is the number of clusters, B_k is the covariance matrix between clusters, W_k is the covariance matrix

Model	Score	Model	Score
GRU	53.60	GRASP+GRU	94.67
RETAIN	36.80	GRASP+RETAIN	52.98
T-LSTM	75.34	GRASP+T-LSTM	116.40
TimeNet$_*$	304.06	GRASP+TimeNet$_*$	457.03
ConCare$_*$	468.76	GRASP+ConCare$_*$	575.72

Table 3: Calinski-Harabasz scores

within clusters, and tr() is the trace of matrix.

We randomly sample 20% of the patients as the test set and use the rest as the training set. We tried several k values for K-means and can observe six main clusters. Therefore we report the clustering C-H score when k = 6. The results are in Table 3. We can see that GRASP helps the backbone models achieve higher C-H scores, which shows its ability to intensify intra-group compactness and inter-group separability.

Next, we conduct **qualitative** cohort studies for the two tasks, mortality prediction and disease diagnosis (i.e., another label for the same dataset, which is defined as a binary classification task of judging whether the patient is diagnosed with diabetes), on CKD dataset to mining medical findings from different perspectives.

Cohort Study on Mortality Prediction. We randomly sample 20% of the patients as the test set and utilize the rest as the training set to perform the mortality prediction task. In a more concise way, the model we use is GRASP with GRU as the backbone. The learned representations are clustered by K-Means (Jain 2010), and statistical analyses are conducted to assess the reasoning of the model. We report the result when cluster number = 6. The result is shown in Table 4. Each row shows the index of the cohort, the death (positive) rate of that cohort, and the distinctive features, respectively. The features are represented using their abbreviations, and the features used in the experiment and their full names are listed in the Appendix. The distinctive features of each cohort are defined as the key distinguishable features to interpret the difference between the cohorts and extract medical findings for the insight. The T-test is used to identify distinctive features. We find that there are 5 to 7 significant features in each cluster, and the top 5 significant features ranked by p-value[4] are reported in Table 4.

Six cohorts of three different types can be observed: low-risk (i.e., cohort # 0 and # 1), medium-risk (i.e., cohort # 2 and # 3) and high-risk (i.e., cohort # 4 and # 5). The death rates of the cohorts are distinct, which shows the learned representations distribute discriminatively. Furthermore, the distinctive features of each cohort are different, especially for the two high-risk cohorts. In Cohort # 4, Serum creatinine (Scr), Weight, Urea, and Appetite are identified as the distinctive features. Those features are important indicators for nutritional status (Carrero 2009; Di Iorio et al. 2018; Gama-Axelsson et al. 2012). Thus, these features reflect the health status from a long-term perspective and are corresponding to the chronic causes of death.

Index	Posi-Rate	Distinctive Features
# 0	0.04	DBP, Cl, SBP, Weight, Glucose
# 1	0.12	Albumin, Hb, Urea, hs-CRP, Scr
# 2	0.39	Scr, Glucose, K, Albumin, Hb
# 3	0.48	Urea, Scr, DBP, Appetite, Weight
# 4	0.74	Scr, Weight, Urea, Appetite, Cl
# 5	0.82	SBP, Albumin, hs-CRP, DBP, Scr

Table 4: Statistics of each cohort w.r.t. mortality prediction

Index	Posi-Rate	Distinctive Features
# 0	0.04	Weight, P, Urea, Glucose, Scr
# 1	0.07	Glu., Scr, Albumin, Appetite, Hb
# 2	0.46	DBP, Scr, Albumin, SBP, K
# 3	0.53	Urea, Glu., DBP, P, Albumin
# 4	0.82	Weight, SBP, Appetite, K, Glu.
# 5	0.90	Glu., Scr, Albumin, DBP, Appetite

Table 5: Statistics of each cohort w.r.t disease diagnosis

In contrast, in the other high-risk cohort, Cohort # 5, Systolic blood pressure (SBP), Albumin, hs-CRP, and DBP are identified as the distinctive features. SBP, hs-CRP, and DBP can reflect the acute changes in health status (Wang et al. 2013; Sarnak et al. 2002), and Albumin is a key feature to evaluate the fundamental health condition (Bal et al. 2013). Specifically, DBP and SBP are essential indicators for heart diseases such as cardiovascular (Kannel 1999) and patients with high hs-CRP are likely to have infections (Aziz et al. 2003), which are corresponding to the acute causes of death.

Cohort Study on Disease Diagnosis. Next, we change the target to the disease diagnosis task on the CKD dataset. We perform the same operation and get cohorts with regard to the diabetes diagnosis. The result is shown in Table 5. We can see that the diagnosed rates of those cohorts are also distinguishable and the distinctive features of each cohort are different. Moreover, the top distinctive features are remarkably different from the ones on mortality prediction. Glucose, of course, has a direct relation with diabetes (Group 2008), and Weight can also reflect the degree of diabetes in a more indirect way (Group et al. 2009). The two high diagnosis rate cohorts (# 4 and # 5) are differentiated by the distinctive features. The above studies show GRASP can distinguish different kinds of similar patients to form cohorts.

Conclusions

In this work, we propose GRASP to boost healthcare models by incorporating auxiliary information from similar patients. It discovers the patients with similar health status, aggregates them into cohorts, and extracts guidance representation. The guidance and the patient information are adaptively fused to depict the health status comprehensively. GRASP demonstrates significant performance improvement, provides medical findings on different COD, and discovers reasonable patient cohorts. The findings are in accord with experts and literature. We hope GRASP can help physicians analyze the patients to prevent the adverse outcome.

[4] The p-values of the reported features are all small than 0.01.

Acknowledgments

This work is supported by the National Science and Technology Major Project (No. 2018ZX10201002) and the Project 2019BD005 supported by PKU-Baidu fund.

References

Aziz, N.; Fahey, J. L.; Detels, R.; and Butch, A. W. 2003. Analytical performance of a highly sensitive C-reactive protein-based immunoassay and the effects of laboratory variables on levels of protein in blood. *Clin. Diagn. Lab. Immunol.* 10(4): 652–657.

Bal, W.; Sokołowska, M.; Kurowska, E.; and Faller, P. 2013. Binding of transition metal ions to albumin: sites, affinities and rates. *Biochimica et Biophysica Acta (BBA)-General Subjects* 1830(12): 5444–5455.

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.

Caliński, T.; and JA, H. 1974. A Dendrite Method for Cluster Analysis. *Communications in Statistics - Theory and Methods* 3: 1–27.

Carrero, J. J. 2009. Identification of patients with eating disorders: clinical and biochemical signs of appetite loss in dialysis patients. *Journal of Renal Nutrition* 19(1): 10–15.

Choi, E.; Bahadori, M. T.; Song, L.; Stewart, W. F.; and Sun, J. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 787–795. ACM.

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.

Chu, X.; Lin, Y.; Wang, Y.; Wang, L.; Wang, J.; and Gao, J. 2019. Mlrda: A multi-task semi-supervised learning framework for drug-drug interaction prediction. In *28th International Joint Conference on Artificial Intelligence*, 4518–4524.

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.

Derogatis, L. R.; Abeloff, M. D.; and Melisaratos, N. 1979. Psychological coping mechanisms and survival time in metastatic breast cancer. *Jama* 242(14): 1504–1508.

Di Iorio, B. R.; Marzocco, S.; Bellasi, A.; De Simone, E.; Dal Piaz, F.; Rocchetti, M. T.; Cosola, C.; Di Micco, L.; and Gesualdo, L. 2018. Nutritional therapy reduces protein carbamylation through urea lowering in chronic kidney disease. *Nephrology Dialysis Transplantation* 33(5): 804–813.

ESTANOL, B. V.; and M. MARIN, O. S. 1975. Cardiac arrhythmias and sudden death in subarachnoid hemorrhage. *Stroke* 6(4): 382–386.

Fried, L. F.; Bernardini, J.; Johnston, J. R.; and Piraino, B. 1996. Peritonitis influences mortality in peritoneal dialysis patients. *Journal of the American Society of Nephrology* 7(10): 2176–2182.

Gama-Axelsson, T.; Heimbürger, O.; Stenvinkel, P.; Bárány, P.; Lindholm, B.; and Qureshi, A. R. 2012. Serum albumin as predictor of nutritional status in patients with ESRD. *Clinical Journal of the American Society of Nephrology* 7(9): 1446–1453.

Gao, J.; Wang, X.; Wang, Y.; Yang, Z.; Gao, J.; Wang, J.; Tang, W.; and Xie, X. 2019. Camp: Co-attention memory networks for diagnosis prediction in healthcare. In *ICDM*, 1036–1041.

Group, C. C. R. 2008. Effects of intensive glucose lowering in type 2 diabetes. *New England journal of medicine* 358(24): 2545–2559.

Group, D. P. P. R.; et al. 2009. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *The Lancet* 374(9702): 1677–1686.

Gumbel, E. J. 1954. Statistical theory of extreme values and some practical applications. *NBS Applied Mathematics Series* 33.

Gupta, P.; Malhotra, P.; Vig, L.; and Shroff, G. 2018. Using Features from Pre-trained TimeNet for Clinical Predictions. In *The 3rd International Workshop on Knowledge Discovery in Healthcare Data at IJCAI*.

Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; et al. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet* 395(10223): 497–506.

Jain, A. K. 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters* 31(8): 651–666.

Kannel, W. B. 1999. Historic perspectives on the relative contributions of diastolic and systolic blood pressure elevation to cardiovascular risk profile. *American heart journal* 138(3): S205–S210.

Kannel, W. B.; Kannel, C.; Paffenbarger Jr, R. S.; and Cupples, L. A. 1987. Heart rate and cardiovascular mortality: the Framingham Study. *American heart journal* 113(6): 1489–1494.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* .

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, F.; You, Q.; Xiao, H.; Chitta, R.; Zhou, J.; and Gao, J. 2018b. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 743–752. ACM.

Ma, L.; Gao, J.; Wang, Y.; Zhang, C.; Wang, J.; Ruan, W.; Tang, W.; Gao, X.; and Ma, X. 2020a. AdaCare: AdaCare: Explainable Clinical Health Status Representation Learning via Scale-Adaptive Feature Extraction and Recalibration. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Ma, L.; Zhang, C.; Wang, Y.; Ruan, W.; Wang, J.; Tang, W.; Ma, X.; Gao, X.; and Gao, J. 2020b. 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.

Maddison, C. J.; Tarlow, D.; and Minka, T. 2014. A* sampling. In *Advances in Neural Information Processing Systems*, 3086–3094.

Prentice, R. L.; and Gloeckler, L. A. 1978. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* 57–67.

Reyna, M. A.; Josef, C. S.; Jeter, R.; Shashikumar, S. P.; Westover, M. B.; Nemati, S.; Clifford, G. D.; and Sharma, A. 2019. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Critical Care Medicine* .

Sarnak, M. J.; Poindexter, A.; Wang, S.-R.; Beck, G. J.; Kusek, J. W.; Marcovina, S. M.; Greene, T.; and Levey, A. S. 2002. Serum C-reactive protein and leptin as predictors of kidney disease progression in the Modification of Diet in Renal Disease Study. *Kidney international* 62(6): 2208–2215.

Suo, Q.; Ma, F.; Yuan, Y.; Huai, M.; Zhong, W.; Gao, J.; and Zhang, A. 2018. Deep patient similarity learning for personalized healthcare. *IEEE transactions on nanobioscience* 17(3): 219–227.

Suresh, H.; Gong, J. J.; and Guttag, J. 2018. Learning Tasks for Multitask Learning: Heterogenous Patient Populations in the ICU. *arXiv preprint arXiv:1806.02878* .

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Tan, Q.; Ye, M.; Yang, B.; Liu, S.; Ma, A. J.; Yip, T. C.-F.; Wong, G. L.-H.; and Yuen, P. 2020. DATA-GRU: Dual-Attention Time-Aware Gated Recurrent Unit for Irregular Multivariate Time Series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 930–937.

Wang, C.; Zhang, J.; Liu, X.; Li, C.; Ye, Z.; Peng, H.; Chen, Z.; and Lou, T. 2013. Reversed dipper blood-pressure pattern is closely related to severe renal and cardiovascular damage in patients with chronic kidney disease. *PloS one* 8(2).

Xu, Y.; Biswal, S.; Deshpande, S. R.; Maher, K. O.; and Sun, J. 2018. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, 2565–2573.

Zhu, Z.; Yin, C.; Qian, B.; Cheng, Y.; Wei, J.; and Wang, F. 2016. Measuring patient similarities via a deep architecture with medical concept embedding. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 749–758. IEEE.