Chronic Disease Management with Personalized Lab Test Response Prediction

Suman Bhoi¹, Mong Li Lee¹, Wynne Hsu¹, Hao Sen Andrew Fang² and Ngiap Chuan Tan²

¹National University of Singapore
²SingHealth Polyclinics

sumanbhoi@u.nus.edu, {leeml, whsu}@comp.nus.edu.sg, {andrew.fang.h.s, tan.ngiap.chuan}@singhealth.com.sg

Abstract

Chronic disease management involves frequent administration of invasive lab procedures in order for clinicians to determine the best course of treatment regimes for these patients. However, patients are often put off by these invasive lab procedures and do not follow the appointment schedules. This has resulted in poor management of their chronic conditions leading to unnecessary disease complications. An AI system that is able to personalize the prediction of individual patient lab test responses will enable clinicians to titrate the medications to achieve the desired therapeutic outcome. Accurate prediction of lab test response is a challenge because these patients typically have co-morbidities and their treatments might influence the target lab test response. To address this, we model the complex interactions among different medications, diseases, lab test response, and fine-grained dosage information to learn a strong patient representation. Together with information from similar patients and external knowledge such as drug-lab interactions and diagnosis-lab interaction, we design a system called KALP to perform personalized prediction of patients’ response for a target lab result and identify the top influencing factors for the prediction. Experiment results on real world datasets demonstrate the effectiveness of KALP in reducing prediction errors by a significant margin. Case studies show that the identified factors are consistent with clinicians’ understanding.

1 Introduction

Electronic Health Records (EHR) provide a rich repository of patient related information such as disease diagnosis, lab test results, prescribed medications, etc. over a history of visits. Advances in machine learning have led to the large-scale analytics of EHR for disease inference [Ni et al., 2017], mortality prediction [Tan et al., 2019], personalized medication recommendation [Wang et al., 2019; Shang et al., 2019; Bhoi et al., 2021]. At the same time, there is a growing trend towards using patient analytics for decision support [Oei et al., 2021] to improve patient care and clinical outcomes, particularly in chronic disease management such as hypertension and diabetes. Often, these chronic disease patients are prescribed some standard treatment regime for each chronic condition independently and their conditions are monitored periodically based on lab test results such as HbA1c¹. For patients with co-morbidities, studies have shown that their lab test results are often influenced by other treatments related to their co-morbidities, making it hard to assess the effectiveness of the prescribed treatment [Unnikrishnan et al., 2012].

The ability to predict the target lab test result of a patient, taking into consideration his condition and medications that are prescribed for co-morbidities, would enable the clinician to personalize the treatment regime for a patient. This can help eliminate invasive procedures associated with sample collection for the lab test, contributing to the Sustainable Development Goal of Good Health and Well-being².

Existing research use patient specific information such as demographics and past visit records to predict lab test results [Luo et al., 2016; Kang, 2018]. These works do not consider the impact of medications on the target lab test result. For example, patients with high blood pressure are prescribed medications like Propranolol which is known to increase HbA1c [Dornhorst et al., 1985]. In practice, lab test results are often influenced by drugs and diagnosis. However, these drug-lab interactions and diagnosis-lab interactions are largely ignored by current works on lab response predictions. Another aspect that has been overlooked is to leverage on patient analytics where patients with similar demographics and diagnosis tend to have similar lab test responses.

In this work, we use a transformer encoder to capture the patient specific information, while the information of similar patients is modeled using the modified graph attention network (GATv2). With this, we obtain a strong latent patient representation and incorporate fine-grained dosage information to accurately predict patient response to a target lab test in the presence of medication titrations for chronic disease patient management. This is in contrast to existing works which mainly consider medication type. The contributions of this work are as follows:

¹https://medlineplus.gov/lab-tests/hemoglobin-a1c-hba1c-test/
²Specific target indicators being addressed are 3.4.1 and 3.d.1 under the third SDG.
• We design a personalized deep learning system that combines the sequential information in patient visit history and information from similar patients to generate a strong patient representation.

• We augment the patient representation with external knowledge of drug-lab interactions and diagnosis-lab interactions to model the complex relationships among drugs, co-morbidities, and lab test results.

• We identify the factors that influenced the predicted lab test results, thus providing insights to changes (if any) in the lab results.

• Extensive quantitative and qualitative experiments on the benchmark MIMIC-III [Johnson et al., 2016] EHR and a proprietary outpatient dataset demonstrate the effectiveness of the proposed system to significantly lower the prediction errors by a large margin.

2 Related Works

Lab test prediction has been studied using patient EHR and visit information [Luo et al., 2016; Kang et al., 2015; Kang et al., 2017; Kang, 2018]. Learning a strong latent patient representation is key to the accurate prediction of lab test results. Researchers have used recurrent neural network (RNN), bidirectional RNN, time-aware long-short term memory and memory augmented RNNs to handle long-range dependencies in patient visit history when learning patient representations [Choi et al., 2016; Ma et al., 2017; Baytas et al., 2017; Sukhbaatar et al., 2015].

Transformer encoder representations have been shown to be superior over RNN based approaches in capturing the sequential dependency in EHR for disease prediction [Li et al., 2020], mortality prediction [Darabi et al., 2020], medication recommendation [Prakash et al., 2021]. HiTANet [Luo et al., 2020] introduces a time-aware transformer to obtain a patient representation using local and global attention for risk prediction. BEHRT [Li et al., 2020] adapts a Transformer-based architecture [Kenton and Toutanova, 2019] to learn patient representation using patient’s diagnostic and demographic information to predict future disease occurrence. All these works do not consider information from other similar patients and do not use fine-grained dosage information.

The work in [Lu et al., 2021] models patient similarity and ontology of diseases to learn disease representation while incorporating clinical notes for the task of diagnosis and heart failure prediction. Our work differs from this in that our notion of patient similarity includes demographics, prescribed medications information in addition to disease diagnosis to predict lab test responses.

3 Methodology

Given a sequence of past patient visits, our task is to predict the result of a target lab test for the patient’s current visit. Figure 1 gives an overview of the proposed solution called KALP for Knowledge Augmented Lab test Prediction. There are two key components: (a) learning an effective latent patient representation and (b) augmenting it with knowledge of drug-lab and disease-lab interactions, as well as patients’ historical responses to the lab test.

3.1 Patient Representation

Our proposed patient representation takes into account both the sequential dependency in patient diagnosis and medication information, as well as information from similar patients. Suppose a patient has \( T - 1 \) visits prior to his current visit. We represent the \( i^{th} \) visit of a patient as

\[
\mathbf{x}_i = [d_i, m_i, l_i]
\]

where \( d_i \) is a multi-hot vector depicting the diagnosis, \( m_i \) is a vector of (medication, dosage) pairs, and \( l_i \) denotes the result of the target lab test, \( 1 \leq i \leq T - 1 \).

Sequential Dependency. We utilize two transformers to model the sequential dependency in the patients’ diagnosis and medication information over the visits. Having dual transformers enable us to obtain representation for patients who may have missing diagnosis or medication information. We linearly embed \( d_i \) and \( m_i \) into a low dimensional space using embedding matrices \( E^d \) and \( E^m \) respectively. These linear embeddings are combined with positional embeddings [Vaswani et al., 2017] to obtain the embeddings for diagnosis \( d_i^e \) and medication \( m_i^e \) which are then passed to their respective transformers (see Figure 2).

Each transformer consists of multiple layers with each layer having \( T - 1 \) encoders. Each encoder has a position-wise fully connected feed-forward network and a multi-head self-attention mechanism. The position-wise feed-forward network has two linear transformations with a ReLU layer in between. The multi-head attention employs scaled dot-product attention to obtain the weights on the patient visits given the visit history. Since this multi-head attention can attend to future time steps, to ensure that the model’s predictions are only conditioned on past visits, we apply a triangular mask to the embedding \( d_i^e \) and \( m_i^e \). This mask is the same as the one used in the decoder component. The input to the \( i^{th} \) encoder in the first layer is the concatenation...
the transformers for diagnosis and medications are given by:

\[ h^{d}_{1}, h^{d}_{2}, \ldots, h^{d}_{T-1} = \text{Transformer}(\{d^{e}_{1}, d^{e}_{2}, \ldots, d^{e}_{T-1}\}) \]  

(1)

\[ h^{m}_{1}, h^{m}_{2}, \ldots, h^{m}_{T-1} = \text{Transformer}(\{m^{c}_{1}, m^{c}_{2}, \ldots, m^{c}_{T-1}\}) \]  

(2)

where \( h^{d}_{i} \) and \( h^{m}_{i} \) are the outputs of the \( i^{th} \) encoder in the first layer and the last layer of the respective transformers. Since our goal is to predict the lab test result for the current visit \( T \), we use the outputs \( h^{d}_{T-1} \) and \( h^{m}_{T-1} \) as the encoded sequential diagnosis and medication information.

**Patient Similarity Information.** Patient similarity has been used to improve the accuracy of diagnosis prediction [Jia et al., 2020] and treatment recommendation [Wirbka et al., 2020]. The work in [MacDonald et al., 2010] observes that patients with similar demographics and diagnosis tend to have similar lab test responses. As such, we incorporate information from similar patients to learn a more effective patient representation to improve the lab result prediction accuracy.

We construct a weighted patient similarity graph \( G = (V, E) \) where each node \( n_{i} \in V \) denotes a patient \( i \), and each labelled edge \( (n_{i}, n_{j}, w_{ij}) \in E \) denotes that patient \( i \) is similar to patient \( j \) with a degree of similarity \( w_{ij} \). Here, \( n_{i} \) is based on patient information at their first visit, together with their age, weight, and gender information.

Note that any similarity measure can be used. Here we use cosine similarity between \( n_{i} \) and \( n_{j} \) to compute \( w_{ij} \). We adapt GATv2 [Brody et al., 2021] to learn the node representation of the weighted patient similarity graph \( G \). GATv2 has two layers with \( z \) attention heads in the first layer, and 1 attention head in the second layer as shown in Figure 2. The attention weight between nodes \( n_{j} \) and \( n_{k} \) for the \( b^{th} \) attention head in the first layer is given by:

\[ \alpha_{jk} = \frac{w_{jk} \times \text{GATv2}(E^{b}, [n_{j}][n_{k}])}{\sum_{n_{i} \in S_{j}} w_{ij} \times \text{GATv2}(E^{b}, [n_{j}][n_{i}])} \]  

(3)

where \( \| \) denotes concatenation, \( S_{j} \) is the set of nodes whose similarity with node \( j \) is non-zero, \( E^{b} \) is the embedding matrix of the \( b^{th} \) attention head, and \( w_{jk} \) denotes the edge weight representing the similarity between patients \( j \) and \( k \).

The output from the first layer for a node \( n_{j} \) is given by:

\[ \alpha_{j} = \| \sum_{b=1}^{z} \sigma(\sum_{n_{k} \in S_{j}} \alpha_{jk} E^{b} \cdot n_{k}) \]  

(4)

where \( \sigma \) is the sigmoid function. With this, the output from the second layer for node \( n_{j} \) can be obtained as follows:

\[ u_{j} = \sigma(\sum_{n_{k} \in S_{j}} \alpha_{jk} E \cdot n_{k}) \]  

(5)

where \( \alpha_{jk} \) is calculated by using the average of the attention weight between node \( j \) and \( k \) from the first layer as new edge weights, i.e., \( w_{jk} \) is updated to the average of \( \alpha_{jk} \) for \( 1 \leq b \leq z \), \( E \) is the embedding matrix for the second layer. Combining this \( u_{j} \) with the diagnosis and medication representation of the most recent visit \( h^{d}_{T-1} \) and \( h^{m}_{T-1} \), we obtain the patient representation \( r_{T-1} \):

\[ r_{T-1} = h^{d}_{T-1} + h^{m}_{T-1} + u_{j} \]  

(6)

### 3.2 Knowledge Augmentation

Next, we augment the obtained patient representation with drug-lab interactions and disease-lab interactions, as well as patients’ past responses to the lab test.

**Lab Interaction Information.** We obtain the drug-lab and disease-lab interactions from AACC Effects on Clinical Laboratory Tests database\(^3\), SIDER [Kuhn et al., 2016] and

\(^3\)https://clinfx.wiley.com/aaccweb/aacc/
MEDI [Wei et al., 2013]. Incorporating this information makes the system aware of the impact of these interactions on the lab test values. Since the interactions can be positive (increases the lab test value) or negative (decreases the lab test value), we use two bipartite graphs to represent them: a positive lab interaction graph $G^+ = (X, Y^+)$ and a negative lab interaction graph $G^- = (X, Y^-)$ where each node $v \in X$ denotes either a lab test, medication or diagnosis and a labelled edge $(v_i, v_j) \in Y^+ / Y^-$ denotes positive/negative interaction between nodes $v_i$ and $v_j$. The lab tests constitute one type of nodes, and medications, diagnosis together constitute another type of nodes and there is no edge between nodes of the same type.

Similar to what we have done for the weighted patient similarity graph, we use GATv2 with two layers and two attention heads in the first layer to learn the node representations from the positive and negative lab interaction graphs (see Figure 3). The node representations of $G^+$ and $G^-$ are mapped to a low dimensional space by applying the embedding matrices $E^{b+}$ and $E^{b-}$ respectively in the first layer for the $b$th attention head where $1 \leq b \leq 2$. With this, we obtain the node representations $p$ and $q$ of the target lab test from $G^+$ and $G^-$ respectively. Then the lab interaction vector $c$ is given by:

$$c = p + \lambda q$$

(7)

where $\lambda$ regulates the fusion of positive and negative lab interaction information.

**Past Lab Test Response Information.** Since patients respond to treatments differently, incorporating a patient’s own past lab test responses will guide KALP to generate personalized lab responses. We have a repository of patients' past lab test responses stored in the form of key-value pairs where the key is the patient representation and the value is the lab test response. Given the patient representation $r_{T-1}$, we compute the attention on the previous representations $r_i$, $1 \leq i \leq T - 2$, stored in the repository as follows:

$$\delta_{T-1} = \text{softmax}(r_1 \cdot r_{T-1}, r_2 \cdot r_{T-1}, ..., r_{T-2} \cdot r_{T-1})$$

(8)

With this attention $\delta_{T-1} \in \mathbb{R}^{T-2}$, we obtain the weighted past visit lab response $f_{T-1}$ as follows:

$$f_{T-1} = \sum_{i=1}^{T-2} \delta_{T-1}[i] \cdot l_{i+1}$$

(9)

where $\delta_{T-1}[i]$ depicts the $i$th entry in the attention vector $\delta_{T-1}$, and $l_{i+1}$ is the lab test result at the $(i + 1)$th visit. By concatenating $r_{T-1}$, $c$, and $f_{T-1}$, we obtain the final output vector which is then passed through a linear layer to predict the result of the target lab test:

$$\hat{y}_T = w \times (r_{T-1} \parallel c \parallel f_{T-1})$$

(10)

where $w$ is the gradient vector of the transformation function in the linear layer. Since the prediction of lab test result is a regression task, our objective function is to minimize the square of the error defined as follows:

$$L_{\text{mse}} = \frac{1}{T-1} \sum_{i=2}^{T} (y_i - \hat{y}_i)^2$$

(11)

where $T$ is the total number of visits, $\hat{y}_i$ and $y_i$ are the predicted and ground-truth lab test result for the $i$th visit.

**3.3 Influential Factors**

The design of our system allows us to identify the factors that have influenced the predicted value $\hat{y}_T$ of the target lab test. Let $w = w^r \parallel w^c \parallel w^f$ where $w^r$, $w^c$, and $w^f$ are the sub-vectors of the linear transformation $w$ in Equation 10 corresponding to patient representation $r_{T-1}$, lab interaction $c$, and past lab test results $f_{T-1}$ respectively.

Then the influence of each diagnosis for the past visits is given by:

$$\eta^d = \|_{i=1}^{T-2} a^d_{T-1}[i] \times (w^r \cdot E^d[:,k]) \times d_i[k]$$

(12)

where $a^d_{T-1}$ is the average of attention weights across all the heads and layers of the encoders in the transformer for the
diagnosis information for time $T - 1$, $a^d_{T-1}[i]$ is the $i^{th}$ entry in $a^d_{T-1}$, $E^d$ is the embedding matrix for diagnosis, and $d_i[k]$ is the $k^{th}$ diagnosis in the $i^{th}$ visit.

Similarly, the influence of a patient’s prescribed medication on the predicted lab test result is:

$$\eta^m = ||_{i=1}^{T-2} a^m_{T-1}[i] \times (w^r \cdot E^m[,k]) \times m_i[k]$$  \hspace{1cm} (13)

where $a^m_{T-1}$ is the average of attention weights across all the heads and layers of the encoders in the transformer for the medication information for time $T - 1$, $a^m_{T-1}[i]$ is the $i^{th}$ entry in $a^m_{T-1}$, $E^m$ is the embedding matrix for medication, and $m_i[k]$ is the $k^{th}$ medication in the $i^{th}$ visit.

We obtain the influence of a patient’s similarity with other patients as follows:

$$\eta^s = ||_{k\in S} a^s_j[k] \times w^r \cdot (E^s \cdot n_k)$$  \hspace{1cm} (14)

where $a^s_j$ is the average of attention weights across all heads and layers of GATv2 for node $j$, $a^s_j[k]$ is the $k^{th}$ entry in $a^s_j$, $E^s$ is the average of embedding matrix $E^b$ over all heads for the nodes in the weighted patient similarity graph, and $n_k$ is the one-hot vector of the node for the $k^{th}$ patient.

The influence of both positive and negative lab interactions are given by:

$$\eta^p = ||_{k\in S} a^p_j[k] \times w^e \cdot (E^p \cdot v_k)$$  \hspace{1cm} (15)

$$\eta^n = ||_{k\in S} a^n_j[k] \times w^e \cdot (E^n \cdot v_k)$$  \hspace{1cm} (16)

where $a^p_j$ and $a^n_j$ are the averages of attention weights across all heads and layers of the GATv2 for the lab interaction graphs $G^+$ and $G^-$ respectively for node $j$, $a^p_j[k]$ is the $k^{th}$ entry in $a^p_j$, $E^p$ and $E^n$ are the average of embedding matrices $E^{b+}$ and $E^{b-}$ over all heads in $G^+$ and $G^-$ respectively, and $v_k$ is the one-hot vector of the $k^{th}$ node in the graphs.

Finally, the influence of a patient’s past lab test responses is given by:

$$\eta^l = ||_{i=1}^{T-2} w^f \times \delta_{T-1}[i] \times l_{i+1}$$  \hspace{1cm} (17)

We normalize all these influences and rank them to obtain the top factors that could have influenced the prediction of the target lab test response.

4 Performance Study

We implemented KALP in PyTorch and trained the models on two NVIDIA Titan RTX GPU. We adopt the widely used metrics for evaluation namely, root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The following datasets are used:

MIMIC-III [Johnson et al., 2016]. This is the largest publicly available EHR dataset which contains clinical data for 7870 neonates (infants) and 38,597 adults admitted to ICU between 2001 and 2008, and captures attributes such as lab reports, medications, etc.

PRIVATE. This is a 10-year outpatient proprietary dataset. Compared to the inpatient MIMIC III, the number of diagnosis per patient is fewer in this dataset.

In our experiments, the target lab test is HbA1c. We filter out the patients who have less than two HbA1c results and perform 10 fold cross validation. We compare KALP with the following baselines:

- **Previous Value (PV).** The predicted lab response for the current visit is given by the previous visit.
- **Linear Regression (LR).** This is the least square method based linear regression.
- **Nearest Neighbour (NN).** The lab test response of the most similar patient is used.
- **RNN [Kang, 2018].** This method employs gated recurrent neural networks to predict lab test results.
- **DMNC [Le et al., 2018].** We adapt this dual memory neural computer to predict lab test response taking into account demographics, medication, and diagnosis.
- **HiTANet [Luo et al., 2020].** We adapt HiTANet to use demographics, medication, and diagnosis to predict lab response.
- **BEHRT [Li et al., 2020].** We adapt BEHRT to use medication in addition to diagnosis and demographics for lab response prediction.

Table 1 shows that KALP outperforms all baselines on both MIMIC-III and PRIVATE. The one-way ANOVA [Fisher, 1992] test shows that the improvements are statistically significant with $p$-values $< 0.05$. Compared to DMNC and BEHRT which model only sequential dependency of a patient’s visits over time, our knowledge augmented approach dramatically widens the performance gap. This suggests that incorporating similar patients, lab interactions and past lab test responses is effective in reducing the prediction errors.

Table 2 shows the performance after medication titration. Here, we use a subset of the patients whose medications have dosage changes between visits. All the methods show an increase in the prediction errors compared to Table 1 as it is difficult to predict the lab test results for these patients which may not follow the general trend after titration. Despite this, KALP has the lowest prediction errors demonstrating its applicability in the real world for chronic disease management.

Additional ablation study shows that not using any medication or patient similarity information leads to the highest RMSE, MAE, and MAPE on both datasets.

5 Case Study

Finally, we present a case study from PRIVATE to show how KALP is able to identify the top influential factors that led to its lab result predictions. Figure 4 shows a chronic diabetic patient with HbA1c ranging between 5.4 and 6.5. KALP predicts the HbA1c with an error margin of 0.2 while the state-of-the-art BEHRT has an error margin as large as 0.7 (see Visit 4). On Visit 3 where the ground truth HbA1c is 5.4, KALP predicts the HbA1c to be 5.5 while BEHRT predicts it to be 5.1. The top influential factor for this prediction is the past lab response, highlighting the need to take into consideration patients’ past responses to personalize the prediction. For Visit 4, we again see that KALP is able to predict the rise in the HbA1c value to within 0.1 of the ground truth value.
compared to BEHRT. A closer examination reveals that there is a reduction in the dosage of Glipizide on Visit 3 and KALP has correctly attributed the top influential factor for this prediction to be the reduction in Glipizide dosage.

### 6 Conclusion

In this work, we have described a personalized lab test result prediction approach that learns a strong patient representation incorporating both patient information accumulated over the visits as well as information from similar patients. This representation captures fine-grained dosage information enabling us to adjust the prediction in response to changes in treatment regime, which is often needed in the management of chronic patient care. To the best of our knowledge, KALP is the only system that takes into account dosage information when making lab test predictions. Experimental results on two real-world datasets demonstrate the effectiveness of KALP in providing predictions that are close to the actual lab results. Future work includes extending KALP to incorporate the severity and frequency of lab interactions, while jointly modeling the plethora of multi-modal EHR information.
Acknowledgements
This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-001-2A). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References


