PERSONALIZED LAB TEST RESPONSE PREDICTION WITH KNOWLEDGE AUGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized medical systems are rapidly gaining traction as opposed to "one size fits all" systems. Predicting patients' lab test responses and providing justification for the predictions would help clinicians tailor treatment regimes for patients. This requires modeling the complex interactions among different medications, diseases, and lab tests. We need to learn a strong patient representation that captures both the sequential information accumulated over the visits and information from other similar patients. We model drug-lab interactions and diagnosis-lab interactions as graphs and design a knowledge-augmented approach to predict patients' response for a target lab result. We also take into consideration patients' past lab responses to personalize the prediction. Experiments on real-world datasets demonstrate the effectiveness of the proposed solution in reducing prediction errors by a significant margin. Case studies show that the identified factors for influencing the predicted results 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 HbA1c[1]. However, 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's condition, taking into consideration medications that are prescribed for his/her other conditions, would enable the clinician to personalize the treatment regime for the patient. This can help eliminate invasive procedures associated with sample collection for the lab test.

Existing research use patient specific information such as demographics and past visit records to predict lab test results (Luo et al., 2016; Kang, 2018). The work in Luo et al. (2016) predicts Ferritin lab test result by using patient demographics and the results of other lab tests that have been ordered at the same time as the Ferritin lab test. Kang (2018) examines the task of predicting HbA1c test results and proposes a recurrent neural network (RNN) based architecture to model the sequential dependencies across the past visits. 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. We observe that drug-lab interaction can be positive where the drug increases the lab test value, or negative where the lab test value is decreased by the drug, e.g., HbA1c is increased

[1] https://medlineplus.gov/lab-tests/hemoglobin-a1c-hba1c-test/

by Propranolol and decreased by Metformin (Dornhorst et al., 1985; González-Ortiz et al., 2012). Similarly, diagnosis-lab interaction can be positive or negative, e.g., HbA1c is increased by iron deficiency anemia and decreased by hemolytic anemia (Program, 2013).

In this work, we use a transformer encoder (Vaswani et al., 2017) to capture the patient specific information, while the information of similar patients is modeled using the modified graph attention network (GATv2)(Brody et al., 2021). We combine the outputs from both the transformer encoder and the GATv2 to obtain a strong latent patient representation so as to accurately predict patient response to a target lab test. We incorporate fine-grained dosage information to increase the accuracy of the predictions in the presence of medication titrations for chronic disease patient management. This is in contrast to existing works which mainly consider medication type.

We augment the patient representation with the knowledge of drug-lab interactions, diagnosis-lab interactions, and use graph attention networks to model the positive and negative effects of drugs and diagnosis on the target lab result. This enables us to model the complex relationships between drugs, co-morbidities, and lab test results. We also take into consideration patients' past lab responses to personalize the prediction. Finally, we leverage the attention weights in the proposed solution to identify the top factors that may have influenced the predicted lab test results. This allows clinicians to have insight into the underlying cause for any changes in the lab test result. 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. Case studies reveal the usefulness of identifying the factors that contributed to the predicted lab test results.

2 Related Works

Lab test prediction has been studied for the purpose of clinical decision support. The work in (Luo et al., 2016) employs multiple machine learning algorithms to predict Ferritin test response from patient demographics and other accompanying lab test responses from EHR. (Kang et al., 2015) uses an ensemble of support vector machines to predict HbA1c test results using patient demographics, prescribed medications, and the last known HbA1c value. (Kang et al., 2017) proposes to use a heterogeneous ensemble of classifiers and reduces the prediction error by incorporating an option to reject an instance when the confidence of any of the classifiers is low. These methods rely on only the current patient visit information and ignore the sequential dependency of past patient visits. (Kang, 2018) proposes a recurrent neural network based architecture to model the sequential dependencies across the visits for the task of prediction of HbA1c test results. Given the black-box nature of RNN, this approach is unable to identify the factors that contribute to the lab test result prediction. In contrast, our solution is able to highlight these factors via attention based mechanism.

Learning a strong latent patient representation is key to the accurate prediction of lab test results. The work in (Choi et al., 2016; Ma et al., 2017) use a recurrent neural network (RNN) and a bi-directional RNN respectively to create a patient representation for disease prediction by capturing the sequential dependency in medical codes over time. Similarly, T-LSTM (Baytas et al., 2017) proposes time aware long-short term memory for patient subtyping. DMNC (Le et al., 2018) learns the patient representation by infusing memory augmented neural networks (Sukhbaatar et al., 2015) with RNNs to handle long-range dependencies in patient visit history. Recent works have demonstrated the superiority of transformer encoder representations 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 Transformer (Vaswani et al., 2017; Devlin et al., 2019) based architecture 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.

Recently, (Lu et al., 2021) model patient similarity and incorporate clinical notes for the task of diagnosis and heart failure prediction. This work uses the ontology of diseases along with the patient diagnosis information to jointly learn the representation for patients and diseases. 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.

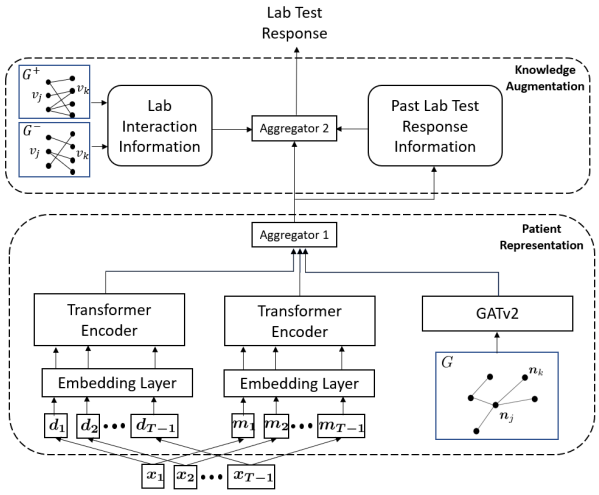


Figure 1: Overview of KALP.

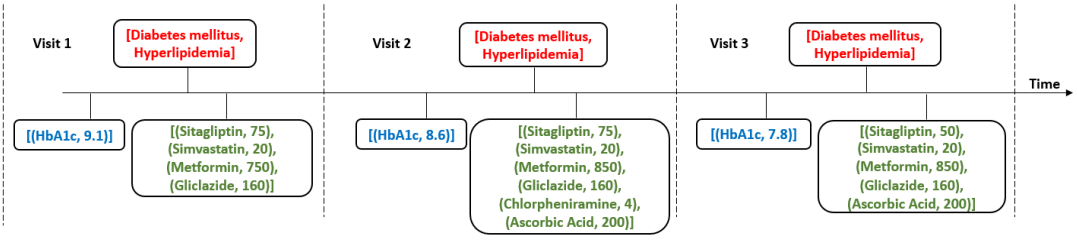


Figure 2: EHR representation of a patient across three visits with HbA1c as the lab test under consideration. Information sources are color-coded with different colors, green indicates medication with dosage, red indicates diagnosis, and blue indicates lab test response.

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-Augment 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 patient's 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

$$\boldsymbol{x}_i = [\boldsymbol{d}_i, \boldsymbol{m}_i, l_i]$$

where $\boldsymbol{d}_i$ is a multi-hot vector depicting the diagnosis, $\boldsymbol{m}_i$ is a vector of (medication, dosage) pairs, and l_i is a continuous variable denoting the results of the target lab test, $1 \leq i \leq T-1$.

We capture the dosage information in the form of a continuous vector where each entry contains the dosage information for the corresponding medication type. All the continuous values are normalized using min-max normalization (Han et al., 2011). Figure 2 illustrates the visit information extracted from a patient's EHR.

Sequential Dependency. We utilize two transformers, one for the diagnosis and the other for medication, 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

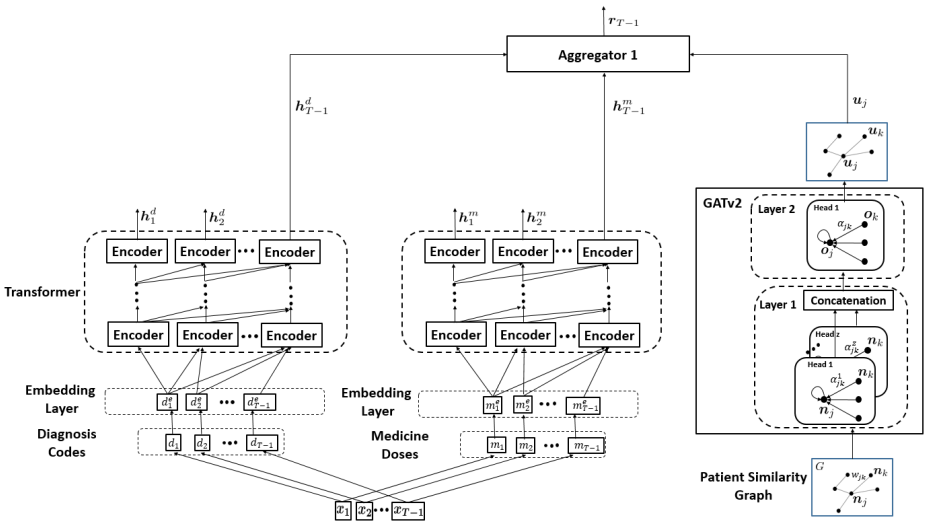


Figure 3: Details of the patient representation component.

have missing diagnosis or medication information. We linearly embed $\boldsymbol{d}_i$ and $\boldsymbol{m}_i$, into a low dimensional space using embedding matrices $\boldsymbol{E}^d$ and $\boldsymbol{E}^m$ respectively. These linear embeddings are then combined with positional embeddings (Vaswani et al., 2017) to obtain the embeddings for diagnosis $\boldsymbol{d}_i^e$ and medication $\boldsymbol{m}_i^e$. The embeddings $\boldsymbol{d}_i^e$ and $\boldsymbol{m}_i^e$ are passed to their respective transformers as shown in Figure 3.

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 $\boldsymbol{d}_i^e$ and $\boldsymbol{m}_i^e$. This mask is the same as the one used in the decoder component of (Vaswani et al., 2017). The input to the i^{th} encoder in the first layer is the concatenation of $\boldsymbol{d}_j^e$, $1 \leq j \leq i$. Inputs to the i^{th} encoder in the subsequent layers is the concatenation of the outputs from the first to the i^{th} encoders of the previous layers.

Similar to (Vaswani et al., 2017), we employ residual connection around the self-attention mechanism and the feed-forward network is followed by layer normalization. We also apply dropout (Srivastava et al., 2014) to avoid over-fitting. The outputs of the transformers for diagnosis and medications are given by:

$$[\boldsymbol{h}_1^d, \boldsymbol{h}_2^d, ..., \boldsymbol{h}_{T-1}^d] = \text{Transformer}\left([\boldsymbol{d}_1^e, \boldsymbol{d}_2^e, ..., \boldsymbol{d}_{T-1}^e]\right) \tag{1}$$

$$[\boldsymbol{h}_1^m, \boldsymbol{h}_2^m, ..., \boldsymbol{h}_{T-1}^m] = \text{Transformer}\left([\boldsymbol{m}_1^e, \boldsymbol{m}_2^e, ..., \boldsymbol{m}_{T-1}^e]\right) \tag{2}$$

where $\boldsymbol{h}_i^d$ and $\boldsymbol{h}_i^m$ are the outputs of i^{th} encoder in 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 $\boldsymbol{h}_{T-1}^d$ and $\boldsymbol{h}_{T-1}^m$ as the encoded sequential diagnosis and medication information.

Patient Similarity Information. The work in (Jia et al., 2020) uses pair-wise patient similarity as a feature and extreme gradient boosting to improve the accuracy of diagnosis prediction. (Wirbka et al., 2020) constructs cohorts of patients who have similar demographics, diagnosis and prescribed medications for treatment recommendation. Here, we observe that patients with similar age, weight, gender, diagnosis, etc. tend to have similar lab test responses. As such, we want to incorporate information from similar patients to learn a more effective patient representation that improves the prediction accuracy of our task.

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}. For this work, we try two popular similarity metrics namely, cosine similarity and Jaccard similarity, and find that cosine similarity has slightly better performance. We use cosine similarity to compute w_{ij} based on patient information at their first visit, together with their age, weight, and gender information. Note that any similarity measure can be used.

Although GAT (Veličković et al., 2018) is a popular choice to obtain node representation from a graph, it represents nodes with limited expressivity (Brody et al., 2021). To mitigate these shortcomings 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 3. The attention weight between nodes n_j and n_k for the b^{th} attention head in the first layer is given by:

$$\alpha_{jk}^{b} = \frac{w_{jk} \times \text{GATv2}(\boldsymbol{E}^{b} \cdot [\boldsymbol{n}_j \| \boldsymbol{n}_k])}{\sum_{\boldsymbol{n}_i \in S_j} w_{ij} \times \text{GATv2}(\boldsymbol{E}^{b} \cdot [\boldsymbol{n}_j \| \boldsymbol{n}_i])} \tag{3}$$

where $\|$ denotes concatenation, S_j is the set of nodes whose similarity with node j is non-zero, GATv2() is the graph attention network operation comprising a single layer feed-forward neural network, followed by a LeakyReLU operation, $\boldsymbol{E}^{b}$ 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:

$$\boldsymbol{o}_j = \|_{b=1}^{z} \sigma \left(\sum_{\boldsymbol{n}_k \in S_j} \alpha_{jk}^{b} \boldsymbol{E}^{b} \cdot \boldsymbol{n}_k \right) \tag{4}$$

where σ is a sigmoid function. With this, the output from the second layer for node n_j can be obtained as follows:

$$\boldsymbol{u}_j = \sigma \left(\sum_{\boldsymbol{n}_k \in \mathcal{S}_j} \alpha_{jk} \boldsymbol{E} \cdot \boldsymbol{o}_k \right) \tag{5}$$

where α_{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 α_{jk}^{b} for $1 \leq b \leq z$, $\boldsymbol{E}$ is the embedding matrix for the second layer. Combining this $\boldsymbol{u}_j$ with the diagnosis and medication representation of the most recent visit $\boldsymbol{h}_{T-1}^{d}$ and $\boldsymbol{h}_{T-1}^{m}$, we obtain the patient representation $\boldsymbol{r}_{T-1}$:

$$\boldsymbol{r}_{T-1} = \boldsymbol{h}_{T-1}^{d} + \boldsymbol{h}_{T-1}^{m} + \boldsymbol{u}_j \tag{6}$$

3.2 KNOWLEDGE AUGMENTATION

Next, we want to augment the patient representation obtained 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 interactions from the *AACC Effects on Clinical Laboratory Tests database*[2], SIDER (Kuhn et al., 2016), MEDI (Wei et al., 2013) and disease-lab interactions from *AACC Effects on Clinical Laboratory Tests database*. 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 graphs to represent them. We have the positive lab interaction graph $G^{+} = (X, Y^{+})$ and negative lab interaction graph $G^{-} = (X, Y^{-})$. Each node $v \in X$ denotes either a lab test, medication or diagnosis. Each labelled edge $(v_i, v_j) \in Y^{+}$ denotes positive interaction between nodes v_i and v_j, while each labelled edge $(v_i, v_j) \in Y^{-}$ denotes negative interaction between nodes v_i and v_j. Note that G^{+} and G^{-} are bipartite graphs where the lab tests constitute one type of nodes, and medications, diagnosis together constitute another type of nodes, in other words, there is no edge between nodes of the same type.

Once again, 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 as we have done for

[2]https://clinfx.wiley.com/aaccweb/aacc/

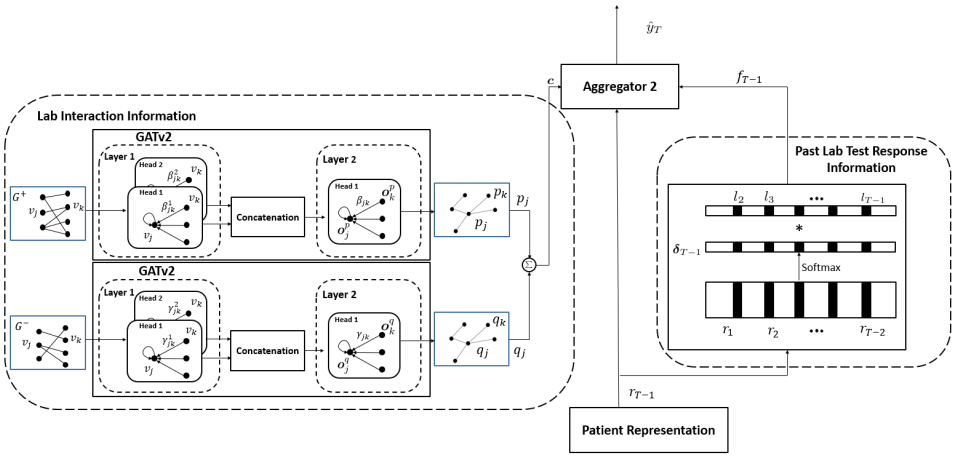


Figure 4: Details of the knowledge augmentation component.

the weighted patient similarity graph (see Figure 4). The node representations of G^+ and G^- are mapped to a low dimensional space by applying the embedding matrices $\boldsymbol{E}^{b+}$ and $\boldsymbol{E}^{b-}$ respectively in the first layer for the b^{th} attention head where $1 \leq b \leq 2$. With this, we can obtain the node representations, $\boldsymbol{p}$ and $\boldsymbol{q}$, of the target lab test from G^+ and G^- respectively. Then the lab interaction vector $\boldsymbol{c}$ is given by:

$$\boldsymbol{c} = \boldsymbol{p} + \lambda \boldsymbol{q} \tag{7}$$

where λ 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. As such, 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 $\boldsymbol{r}_{T-1}$, we compute the attention on the previous representations $\boldsymbol{r}_i$, $1 \leq i \leq T-2$, stored in the repository as follows:

$$\boldsymbol{\delta}_{T-1} = softmax(\boldsymbol{r}_1 \cdot \boldsymbol{r}_{T-1},\ \boldsymbol{r}_2 \cdot \boldsymbol{r}_{T-1}, ..., \boldsymbol{r}_{T-2} \cdot \boldsymbol{r}_{T-1}) \tag{8}$$

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

$$\boldsymbol{f}_{T-1} = \sum_{i=1}^{T-2} \delta_{T-1}[i] \cdot l_{i+1} \tag{9}$$

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

$$\hat{y}_T = \boldsymbol{w} \times (\boldsymbol{r}_{T-1} \parallel \boldsymbol{c} \parallel f_{T-1}) \tag{10}$$

where $\boldsymbol{w}$ 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:

$$\mathcal{L}_{\text{mse}} = \frac{1}{T-1} \sum_{i=2}^{T} (y_i - \hat{y}_i)^2 \tag{11}$$

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 respectively.

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 based on the attention scores of GAT. Let $\boldsymbol{w}^r$, $\boldsymbol{w}^c$, and $\boldsymbol{w}^f$ be the sub-vectors of the linear transformation in Equation 10 corresponding to patient representation $\boldsymbol{r}_{T-1}$,

lab interaction $\boldsymbol{c}$, and past lab test results $\boldsymbol{f}_{T-1}$ respectively. In other words, we have

$$\boldsymbol{w} = \boldsymbol{w}^r \parallel \boldsymbol{w}^c \parallel \boldsymbol{w}^f$$

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

$$\boldsymbol{\eta}^d = \|_{i=1}^{T-2} a_{T-1}^d[i] \times (\boldsymbol{w}^r \cdot \boldsymbol{E}^d[:,k]) \times \boldsymbol{d}_i[k] \tag{12}$$

where $\boldsymbol{a}_{T-1}^d$ is the average of attention weights across all the heads and layers of the encoders in the transformer for the diagnosis information for time $T-1$, $a_{T-1}^d[i]$ is the i^{th} entry in $\boldsymbol{a}_{T-1}^d$, $\boldsymbol{E}^d$ is the embedding matrix for diagnosis, and $\boldsymbol{d}_i[k]$ 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:

$$\boldsymbol{\eta}^m = \|_{i=1}^{T-2} a_{T-1}^m[i] \times (\boldsymbol{w}^r \cdot \boldsymbol{E}^m[:,k]) \times \boldsymbol{m}_i[k] \tag{13}$$

where $\boldsymbol{a}_{T-1}^m$ is the average of attention weights across all the heads and layers of the encoders in the transformer for the medication information for time $T-1$, $a_{T-1}^m[i]$ is the i^{th} entry in $\boldsymbol{a}_{T-1}^m$, $\boldsymbol{E}^m$ is the embedding matrix for medication, $\boldsymbol{m}_i[k]$ is the k^{th} medication in the i^{th} visit. We obtain the influence of a patient's similarity with other patients as follows:

$$\boldsymbol{\eta}^s = \|_{k \in \mathcal{S}_j} a_j^s[k] \times \boldsymbol{w}^r \cdot (\boldsymbol{E}^s \cdot \boldsymbol{n}_k) \tag{14}$$

where $\boldsymbol{a}_j^s$ is the average of attention weights across all heads and layers of GATv2 for node j, $a_j^s[k]$ is the k^{th} entry in $\boldsymbol{a}_j^s$, $\boldsymbol{E}^s$ is the average of embedding matrix $\boldsymbol{E}^b$ over all heads for the nodes in the weighted patient similarity graph, and $\boldsymbol{n}_k$ is the one-hot vector of the node for the k^{th} patient. The influence of both positive and negative lab interactions can be obtained as follows:

$$\boldsymbol{\eta}^p = \|_{k \in \mathcal{S}_j} a_j^p[k] \times \boldsymbol{w}^c \cdot (\boldsymbol{E}^p \cdot \boldsymbol{v}_k) \tag{15}$$

$$\boldsymbol{\eta}^q = \|_{k \in \mathcal{S}_j} \lambda \times a_j^q[k] \times \boldsymbol{w}^c \cdot (\boldsymbol{E}^q \cdot \boldsymbol{v}_k) \tag{16}$$

where $\boldsymbol{a}_j^p$ and $\boldsymbol{a}_j^q$ 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, $a_j^p[k]$ is the k^{th} entry in $\boldsymbol{a}_j^p$ and $a_j^q[k]$ is the k^{th} entry in $\boldsymbol{a}_j^q$, $\boldsymbol{E}^p$ and $\boldsymbol{E}^q$ are the average of embedding matrices $\boldsymbol{E}^{b+}$ and $\boldsymbol{E}^{b-}$ over all heads in G^+ and G^- respectively, and $\boldsymbol{v}_k$ 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:

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

We normalize these influences and rank them to obtain the top factors that could have influenced the prediction of the target lab test response. These scores need not necessarily mean explanability/interpretability in general context.

4 Performance Study

In this section, we evaluate the effectiveness of our proposed KALP in predicting lab test results. KALP[3] is implemented in PyTorch and the models are trained on two NVIDIA Titan RTX GPU. We adopt the widely used metrics for regression namely, root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) in our evaluation (see Appendix A.1). We use the following datasets in our experiments:

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. We divide the datasets into training, validation, and test sets in the ratio of 8:1:1. We report the results in the format $\mu \pm \sigma$ where μ and σ are the mean and standard deviation over the

[3]The code will be available in Github.

Table 1: Results for comparative study.

Methods	MIMIC-III			PRIVATE		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
PV	2.39 ± 0.21	1.91 ± 0.19	14.47 ± 2.51	2.08 ± 0.17	1.95 ± 0.19	13.22 ± 2.18
LR	2.91 ± 0.25	2.52 ± 0.27	18.26 ± 2.76	2.38 ± 0.21	2.05 ± 0.20	15.34 ± 2.45
NN	2.61 ± 0.20	2.18 ± 0.21	16.58 ± 2.57	2.11 ± 0.19	1.94 ± 0.17	14.48 ± 2.33
RNN	2.37 ± 0.18	1.88 ± 0.19	14.04 ± 2.21	2.04 ± 0.17	1.72 ± 0.15	12.53 ± 2.13
DMNC	1.95 ± 0.16	1.67 ± 0.15	13.19 ± 2.08	1.83 ± 0.15	1.55 ± 0.12	10.64 ± 2.01
HiTANet	1.89 ± 0.17	1.51 ± 0.15	12.37 ± 2.11	1.72 ± 0.14	1.43 ± 0.11	9.52 ± 1.93
BEHRT	1.57 ± 0.14	1.38 ± 0.13	11.28 ± 1.95	1.41 ± 0.11	1.24 ± 0.13	8.49 ± 1.82
KALP	$\mathbf{1.15 \pm 0.11}$*	$\mathbf{0.80 \pm 0.10}$*	$\mathbf{6.87 \pm 1.63}$*	$\mathbf{0.85 \pm 0.09}$*	$\mathbf{0.69 \pm 0.07}$*	$\mathbf{3.42 \pm 1.47}$*

* indicates that the result is statistically significant when compared to the second best with p-value < 0.05.

Table 2: Results when there are changes in medication dosage.

Methods	MIMIC-III			PRIVATE		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
PV	4.16 ± 0.23	3.73 ± 0.21	23.11 ± 2.61	3.91 ± 0.24	3.27 ± 0.21	20.52 ± 2.32
LR	4.56 ± 0.31	3.98 ± 0.28	24.03 ± 2.89	4.05 ± 0.29	3.43 ± 0.26	21.04 ± 2.67
NN	4.12 ± 0.28	3.69 ± 0.27	22.45 ± 2.77	3.87 ± 0.24	3.22 ± 0.22	20.26 ± 2.51
RNN	3.62 ± 0.25	3.26 ± 0.26	20.64 ± 2.64	3.34 ± 0.23	3.01 ± 0.21	19.16 ± 2.39
DMNC	3.18 ± 0.22	2.76 ± 0.23	19.01 ± 2.55	2.89 ± 0.21	2.33 ± 0.20	17.02 ± 2.25
HiTANet	3.08 ± 0.19	2.75 ± 0.15	18.41 ± 2.41	2.65 ± 0.18	2.19 ± 0.17	16.18 ± 2.21
BEHRT	2.65 ± 0.18	2.39 ± 0.17	17.70 ± 2.43	2.26 ± 0.16	2.02 ± 0.15	15.02 ± 2.23
KALP	$\mathbf{1.85 \pm 0.13}$*	$\mathbf{1.34 \pm 0.12}$*	$\mathbf{10.28 \pm 2.13}$*	$\mathbf{1.27 \pm 0.10}$*	$\mathbf{1.10 \pm 0.11}$*	$\mathbf{8.18 \pm 1.96}$*

* indicates that the result is statistically significant when compared to the second best with p-value < 0.05.

10-folds. Appendix A.2 gives the statistics of the resulting datasets, pre-processing steps, training and optimization details. Appendix A.3 shows the optimal parameters as determined by sensitivity experiments. We compare KALP with the following methods:

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). The original DMNC only considers diagnosis and procedure. Here we adapt it to predict lab response taking into account demographics, medication, and diagnosis.

HiTANet (Luo et al., 2020). The original HiTANet uses only diagnosis to perform risk prediction. We adapt this to predict lab response by using demographics, medication, and diagnosis.

BEHRT (Li et al., 2020). This BERT-based technique originally takes as input diagnosis and demographics, but we have adapted it to provide lab response prediction by using medication in addition to demographics and diagnosis.

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, we see that 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 lowering the prediction errors. The ablation study in Appendix A.4 shows the effect of different information sources. Using different neural architectures show that KALP continues to give good performance over existing methods (see Appendix A.5).

Table 2 shows the performance of the various methods to predict lab test results after the clinician has prescribed a change in the medication dosage. Here, we use a subset of the patients whose medications have dosage changes between visits. Compared to Table 1, all the methods show an increase in the prediction errors. This is because the lab test result for patients with changes in medication

dosage may not follow the general trend in the EHR and would be difficult to predict their lab test results. We observe that KALP has the lowest prediction errors, demonstrating its applicability in the real world where clinicians often titrate medication dosage for patient management.

5 CASE STUDY

Finally, we present a case study from PRIVATE to demonstrate the performance of KALP and its ability to identify the top influencing factors that led to its lab result predictions. The top influencing factors can be obtained from the attention weights (recall Section 3.3).

Figure 5 shows a chronic diabetic patient with HbA1c ranging between 5.4 and 6.5. KALP is able to predict the HbA1c to an accuracy of 0.2 while the state-of-the-art BEHRT has a prediction error as large as 0.7 (Visit 4). On visit 3, we see that KALP predicted the HbA1c to be 5.5 while BEHRT predicted it to be 5.1. The top influencing factor for this prediction is the past lab response. This highlights that taking into consideration patients' past responses helps in personalizing the prediction. Further for Visit 4, we again see that KALP is able to predict the rise in the HbA1c value 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 attributed the top influencing factor for this prediction to be the reduction in dosage of Glipizide.

To the best of our knowledge, KALP is the only system that takes into account dosage information when making lab test predictions. In practice, clinicians often need to adjust medication treatment to manage patient care. With KALP's ability to model medication dosages, we have demonstrated here a highly accurate prediction for the target lab test result. Appendix A.6 provides additional case studies to demonstrate the importance of modeling other patient information such as medication dosage, patients' past lab responses.

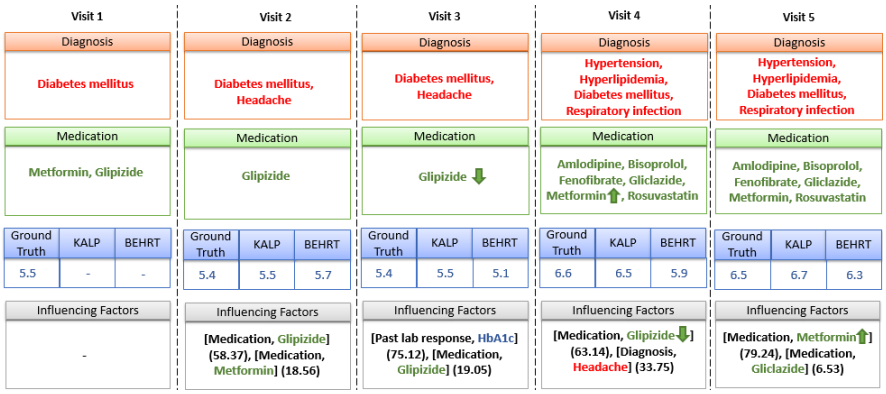


Figure 5: Predicted and ground truth HbA1c. Influencing factors are obtained from KALP. Blue depicts HbA1c values, red depicts diagnosis, and green depicts medication. ⇓ and ⇑ depict decrease and increase of dosage respectively.

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 also 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 increase the prediction accuracy for patients with complex co-morbidities, we have augmented the patient representation with external knowledge on lab interactions and patients' historical responses to the target lab test. 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 information on the severity and frequency of lab interactions. Further, KALP only uses the structured information in the EHR. However, EHR encompasses a plethora of information in the form of doctors' notes, physiological signals, and medical images. One promising direction would be to jointly model these multi-modal sources in KALP.

REFERENCES

Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 65–74, 2017.

Suman Bhoi, Mong Li Lee, Wynne Hsu, Hao Sen Andrew Fang, and Ngiap Chuan Tan. Personalizing medication recommendation with graph-based approach. *Transactions on Information Systems*, 1(1), 2021. doi: 10.1145/3488668.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

Edward Choi, Mohammad Taha Bahadori, Joshua A Kulas, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems*, pp. 3512–3520, 2016.

Sajad Darabi, Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. Taper: Time-aware patient ehr representation. *IEEE journal of biomedical and health informatics*, 24(11):3268–3275, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

Anne Dornhorst, StephenH Powell, and Jack Pensky. Aggravation by propranolol of hyperglycaemic effect of hydrochlorothiazide in type ii diabetics without alteration of insulin secretion. *The Lancet*, 325(8421):123–126, 1985.

Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pp. 66–70. Springer, 1992.

Manuel González-Ortiz, Esperanza Martínez-Abundis, José A Robles-Cervantes, Maria G Ramos-Zavala, Carmelita Barrera-Durán, and Jorge González-Canudas. Effect of metformin glycinate on glycated hemoglobin a1c concentration and insulin sensitivity in drug-naive adult patients with type 2 diabetes mellitus. *Diabetes technology & therapeutics*, 14(12):1140–1144, 2012.

Jiawei Han, Micheline Kamber, and Jian Pei. Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, 5(4):83–124, 2011.

Zheng Jia, Xian Zeng, Huilong Duan, Xudong Lu, and Haomin Li. A patient-similarity-based model for diagnostic prediction. *International journal of medical informatics*, 135:104073, 2020.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Seokho Kang. Personalized prediction of drug efficacy for diabetes treatment via patient-level sequential modeling with neural networks. *Artificial intelligence in medicine*, 85:1–6, 2018.

Seokho Kang, Pilsung Kang, Taehoon Ko, Sungzoon Cho, Su-jin Rhee, and Kyung-Sang Yu. An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. *Expert Systems with Applications*, 42(9):4265–4273, 2015.

Seokho Kang, Sungzoon Cho, Su-jin Rhee, and Kyung-Sang Yu. Reliable prediction of anti-diabetic drug failure using a reject option. *Pattern Analysis and Applications*, 20(3):883–891, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR*, 2015.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079, 2016.

Hung Le, Truyen Tran, and Svetha Venkatesh. Dual memory neural computer for asynchronous two-view sequential learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1637–1645, 2018.

Yikuan Li, Shishir Rao, Jose Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12, 2020.

Chang Lu, Chandan K. Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 3529–3535. ijcai.org, 2021. doi: 10.24963/ijcai.2021/486. URL https://doi.org/10.24963/ijcai.2021/486.

Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 647–656, 2020.

Yuan Luo, Peter Szolovits, Anand S Dighe, and Jason M Baron. Using machine learning to predict laboratory test results. *American journal of clinical pathology*, 145(6):778–788, 2016.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 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*, pp. 1903–1911, 2017.

Jingchao Ni, Hongliang Fei, Wei Fan, and Xiang Zhang. Cross-network clustering and cluster ranking for medical diagnosis. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pp. 163–166. IEEE, 2017.

Ronald Wihal Oei, Hao Sen Andrew Fang, Wei-Ying Tan, Wynne Hsu, Mong-Li Lee, and Ngiap-Chuan Tan. Using domain knowledge and data-driven insights for patient similarity analytics. *Journal of Personalized Medicine*, 11(8):699, Jul 2021. ISSN 2075-4426. doi: 10.3390/jpm11080699.

PKS Prakash, Srinivas Chilukuri, Nikhil Ranade, and Shankar Viswanathan. Rarebert: Transformer architecture for rare disease patient identification using administrative claims. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 453–460, 2021.

National Glycohemoglobin Standardization Program. Factors that interfere with hba1c test results, 2013. URL http://www.ngsp.org/factors.asp.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1126–1133, 2019.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. *Advances in Neural Information Processing Systems*, 2015:2440–2448, 2015.

Qingxiong Tan, Andy Jinhua Ma, Mang Ye, Baoyao Yang, Huiqi Deng, Vincent Wai-Sun Wong, Yee-Kit Tse, Terry Cheuk-Fung Yip, Grace Lai-Hung Wong, Jessica Yuet-Ling Ching, et al. Ua-crnn: Uncertainty-aware convolutional recurrent neural network for mortality risk prediction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 109–118, 2019.

H Thimas, CE Cormen, RL Leiserson, and RC Stein. Introduction to algorithms, 2009.

Ranjit Unnikrishnan, Ranjit Mohan Anjana, and Viswanathan Mohan. Drugs affecting hba1c levels. *Indian journal of endocrinology and metabolism*, 16(4):528, 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *6th International Conference on Learning Representations, ICLR*, 2018.

Shanshan Wang, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke. Order-free medicine combination prediction with graph convolutional reinforcement learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1623–1632, 2019.

Wei-Qi Wei, Robert M Cronin, Hua Xu, Thomas A Lasko, Lisa Bastarache, and Joshua C Denny. Development and evaluation of an ensemble resource linking medications to their indications. *Journal of the American Medical Informatics Association*, 20(5):954–961, 2013.

Lucas Wirbka, Walter E Haefeli, and Andreas D Meid. A framework to build similarity-based cohorts for personalized treatment advice–a standardized, but flexible workflow with the r package simbaco. *PloS one*, 15(5):e0233686, 2020.

A APPENDIX

A.1 EVALUATION METRICS

Suppose T is the total number of visits, $\hat{y}_i$ and y_i are the predicted and ground-truth lab test results for the i^{th} visit respectively. The formula of the evaluation metrics used are as follows:

$$\text{RMSE} = \sqrt{\frac{1}{T-1}\sum_{i=2}^{T}(y_i - \hat{y}_i)^2} \tag{18}$$

$$\text{MAE} = \frac{1}{T-1}\sum_{i=2}^{T}|y_i - \hat{y}_i| \tag{19}$$

$$\text{MAPE} = \frac{100}{T-1}\sum_{i=2}^{T}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{20}$$

A.2 DATASET STATISTICS, PRE-PROCESSING, TRAINING AND OPTIMIZATION

Table 3 and Table 4 give the statistics and the patient characteristics of the experiment datasets respectively.

Table 3: Characteristics of datasets.

Attribute	**MIMIC-III**	**PRIVATE**
Number of patients	6412	6312
Number of diagnosis	1919	139
Number of medications	2941	45
Average number of visits per patient	2.67	13.06
Average number of diagnosis per visit	13.47	5.90
Average number of medications per visit	32	4.07
Maximum number of diagnosis per visit	39	27
Maximum number of medications per visit	148	12

For both MIMIC-III and PRIVATE we use the ICD-9 coding system for diagnosis and the generic names for medications. Some of the important pre-processing steps are discussed next. Since the dosages are reported in different units we convert all the dosage values to correspond to either milligram or milliliter. For this study, we distinguish unique visits based on the unique hospital admission time (given by unique hospital ID in the datasets). Furthermore, in case of multiple dosage prescriptions of the same medicine and multiple lab test results within the same visit we consider their latest entries. We also filter out the patients who have less than two instances of the target lab test result.

The embedding size for our model is fixed at 128 and training is done using Adam (Kingma & Ba, 2015). The learning rate is 0.0002, and the best performing model is chosen on the validation set after 50 epochs.

For MIMIC-III, we apply a dropout of 0.4 on the input embedding layer. The transformer encoder consists of four layers with 4 attention heads each. The graph attention network (GATv2) used to

Table 4: Patient characteristics of both datasets.

Variable	**Type**	**MIMIC-III**	**PRIVATE**
Age	Discrete	65.36 (13.27)	58.45 (15.67)
Gender	Categorical	4136 (M), 2276 (F)	3324 (M), 2988 (F)
Weight	Continuous	66.23 (15.25)	71.45 (14.12)
HbA1c	Continuous	9.2 (3.9)	10.5 (4.2)

model patient similarity consists of two layers with 2 attention head in the first layer and 1 attention head in the second layer. The parameter value is set to be $\lambda = 0.49$ based on the validation set. The learning rate is 0.0001, and the best performing model is chosen based on the performance on the validation set after 50 epochs. The hyper-parameters for all the baselines were chosen on the validation set. For Recurrent Neural Network we use GRU as the recurrent unit with a hidden dimension of 64. The transformer component of HiTANet has 2 layers, 6 attention heads, intermediate layer size of 1024, hidden size of 256 while other parameters remain the same as the original work. The time interval is set as 1 for HiTANet. BEHRT with an architecture of 4 layers, 6 attention heads, intermediate layer size of 512, and hidden size of 288 provided optimal performance. The word and memory size for DMNC model is 64 and 16 as used in the original work itself.

For PRIVATE, the dropout is 0.6 and the transformer encoder consists of two layers with 4 attention heads each. The graph attention network (GATv2) used to model patient similarity consists of two layers with 4 attention heads in the first layer and 1 attention head in the second layer. The parameter value is $\lambda = 0.35$. The recurrent unit used for Recurrent Neural Network was GRU with a hidden dimension of 64. The transformer component of HiTANet has 1 layer, 2 attention heads, intermediate layer size of 1024, hidden size of 256 while other parameters remain the same as the original work. The time interval is set as 1 for HiTANet. BEHRT with an architecture of 2 layers, 4 attention heads, intermediate layer size of 512, and hidden size of 288 provided optimal performance. The word and memory size for DMNC model is 64 and 32.

A.3 Sensitivity Experiments

We examine the effect of embedding dimension for the inputs, number of layers and the number of attention heads in the transformer encoders on the performance of KALP.

Figure 6(a) shows the results as we vary the input embedding dimension from 8 to 256. We see that the smallest RMSE is achieved when the embedding dimension is 128 for both MIMIC-III and PRIVATE datasets.

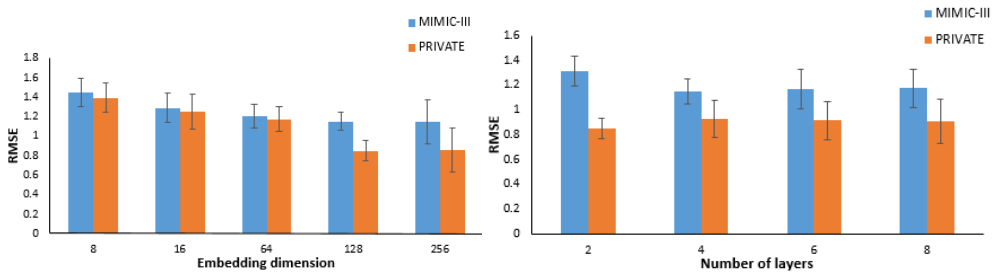
Figure 6(b) shows how KALP performs when we vary the number of transformer layers. We observe that having 4 layers gives the best performance in MIMIC-III whereas 2 layers is sufficient for the PRIVATE dataset. Figure 6(c) shows that the best results is obtained when the number of attention heads in the transformer is 4 for both the datasets.

We also vary the number of attention heads in the first layer of GATv2 used to model the patient similarity graph. Figure 6(d) shows that having 2 attention heads gives the smallest RMSE on MIMIC-III dataset while having 4 attention heads leads to the best performance for the PRIVATE dataset.

A.4 Ablation Study

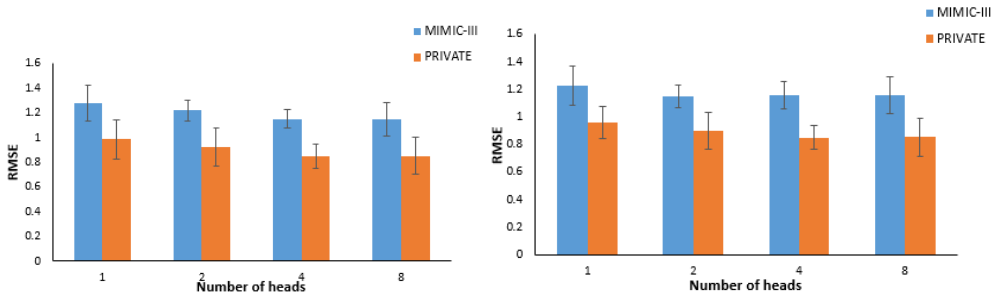
We perform an ablation study to understand the effect of knowledge from different information sources on the performance of KALP. We implement the following variants of KALP:

- **KALP w/o medication.** This variant uses all the information sources except the prescribed medications and their dosages.
- **KALP w/o similar patients.** This variant does not include information from similar patients in the patient representation.
- **KALP w/o lab interactions.** This model does not use any information from the lab interactions. This depicts the model variant not using any external information.
- **KALP w/o positive lab interaction.** Here, we do not use the positive lab interaction information.
- **KALP w/o negation lab interaction.** This model does not use the negative lab interaction information.
- **KALP w/o past lab test results.** Here, we do not use information from the patient's past lab test results.
- **KALP w/o knowledge augmentation.** In this final variation, we do not use information from the patient's past lab test results as well as the lab interaction information.



(a) Dimension of input embedding layer

(b) Layers of transformer encoder



(c) Attention head of transformer encoder

(d) Attention head of GATv2

Figure 6: Effect of embedding dimension and attention heads on PREMIER.

Table 5: Ablation Study

Methods	MIMIC-III			PRIVATE		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
KALP w/o medication	1.55	1.28	11.01	1.21	1.08	8.07
KALP w/o similar patients	1.49	1.22	8.39	1.15	1.04	7.75
KALP w/o lab interactions	1.35	1.13	8.24	1.11	0.91	7.50
KALP w/o positive lab interactions	1.22	0.98	7.55	0.91	0.79	6.66
KALP w/o negative lab interactions	1.28	1.02	7.63	0.99	0.87	7.43
KALP w/o past lab results	1.42	1.18	8.42	1.15	1.01	7.59
KALP w/o knowledge augmentation	1.45	1.20	8.40	1.16	1.03	7.62
BEHRT	1.57	1.38	11.28	1.41	1.24	8.49
KALP	**1.15**	**0.80**	**6.87**	**0.85**	**0.69**	**3.42**

Table 5 shows the performance of these variants on MIMIC-III and PRIVATE datasets. We observe that the best performance is achieved when all the information sources are utilized. Not using any medication or patient similarity information leads to the highest RMSE, MAE, and MAPE on both datasets. This indicates that having a patient representation that incorporates the patient's prescribed medications and information from similar patients can boost the accuracy of lab test result predictions. We observe that the performance of KALP without the knowledge augmentation component still outperforms the state-of-the-art technique, BEHRT suggesting that the patient representation learnt by KALP is indeed effective.

A.5 Architectural Variants of KALP

We also carry out experiments to understand the impact of architectural changes on the performance of KALP. The following variants of KALP are implemented:

- **KALP**$_{GAT}$**.** We use graph attention network, GAT (Veličković et al., 2018) to model the patient similarity and the lab interaction information. The rest of the architecture is the same as that of KALP.

- **KALP**$_{sparse}$**.** We sparsify the patient similarity graph by retaining edges whose similarity scores are above some threshold. Here we set the threshold to 0.95.
- **KALP**$_{MST}$**.** We sparsify the patient similarity graph by using a minimum spanning tree, MST (Thimas et al., 2009) to minimize the inverse of the edge weights (i.e. similarity scores) .

Table 6 shows that KALP has the best performance compared to KALP$_{GAT}$, KALP$_{sparse}$, and KALP$_{MST}$.

Table 6: Comparison of Architectural Variants of KALP.

Methods	MIMIC-III			PRIVATE		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
KALP$_{GAT}$	1.35	1.11	7.21	1.09	0.91	5.17
KALP$_{sparse}$	1.27	1.02	7.09	0.98	0.85	4.56
KALP$_{MST}$	1.19	0.91	6.95	0.89	0.76	3.72
KALP	**1.15**	**0.80**	**6.87**	**0.85**	**0.69**	**3.42**

A.6 Additional Case Studies

Recall that PRIVATE is an outpatient dataset with patients having chronic diseases and co-morbidities. These patients have regular follow-up visits to monitor and manage their conditions. This case study demonstrates the importance of capturing medication dosage and patients' past lab response.

Figure 7 shows a Patient A from the PRIVATE dataset who has a HbA1c test result of 11.4 during his first visit to the clinic and is diagnosed with Acute upper respiratory infection, Hyperlipidemia, Diabetes mellitus. We see that KALP is able to accurately predict the HbA1c values within a 0.1 margin, whereas BEHRT's margin of error tends to be larger, particularly for the prediction at Visit 5. This is because prior to Visit 5, there is a downward titration of Metformin and the reduced dosage may lead to an increase in the HbA1c value. KALP is able to infer the rise in HbA1c values as it has captured the medication dosage in its patient representation, whereas BEHRT does not capture such dosage information. This is confirmed by the top influencing factors for Visit 5 where the top factor is indicated as the reduced dosage in Metformin.

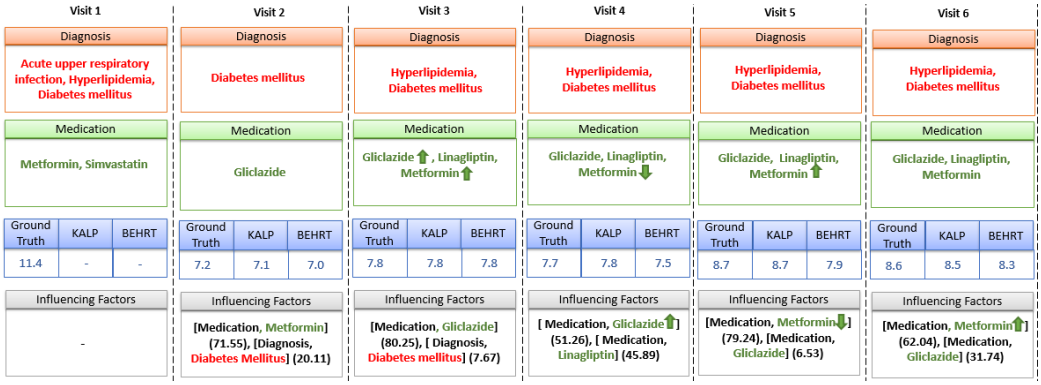


Figure 7: Predicted and ground truth HbA1c results for Patient A. Influencing factors are obtained from KALP.

A.6.1 Samples from MIMIC-III.

The patients in MIMIC-III are in ICU with acute conditions and complex co-morbidities. These case studies demonstrate the importance of taking into account knowledge of the lab interactions.

Figure 8 shows a Patient B from the MIMIC-III with a HbA1c value of 5.5 at his first visit and is diagnosed with multiple diseases such as congestive heart failure, diabetes, hypertension, etc. He is

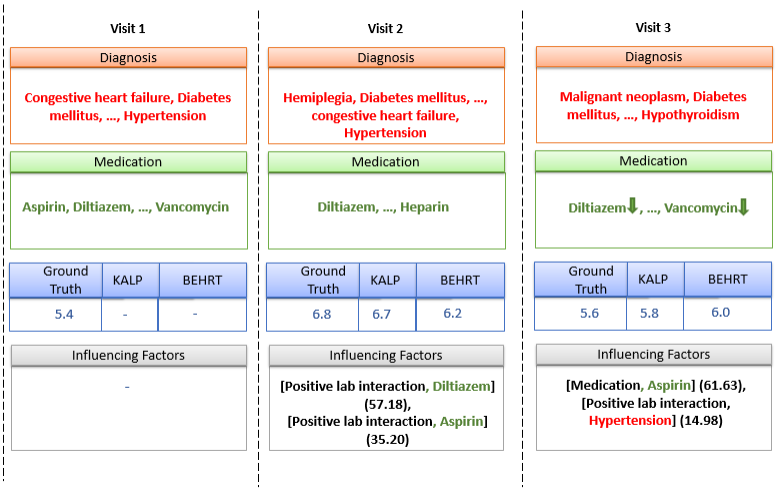


Figure 8: Predicted and ground truth HbA1c for Patient B. Influencing factors are obtained from KALP.

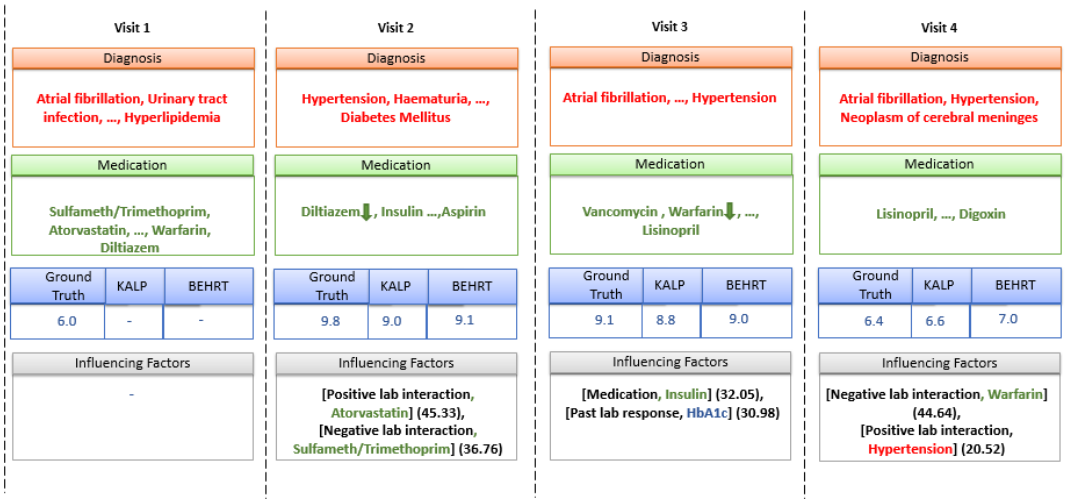


Figure 9: Predicted and ground truth HbA1c results for Patient C. Influencing factors are obtained from KALP.

prescribed with various medications including Asprin and Diltiazem. Both Diltiazem and Aspirin are known to increase the HbA1c value. With the augmented lab interaction knowledge, KALP is able to accurately predict the HbA1c value within 0.1 of the actual lab result, and has correctly identified the top influencing factors as the positive lab interaction of Diltiazem and Aspirin on HbA1c values in Visit 2.

Figure 9 shows the details of Patient C and the HbA1c predictions by KALP and BEHRT along with the influencing factors identified by KALP. We see that the patient has a HbA1c value of 6.0 at Visit 1 and is diagnosed with urinary tract infection (UTI), atrial fibrillation, hyperlipidemia, etc. Medications like Trimethoprim, Atorvastatin, etc. are prescribed to this patient.

At Visit 2, KALP predicted an increase in the HbA1c value and identified the top two influencing factors as the positive lab interaction of Atorvastatin and the negative lab interaction of Trimethoprim on HbA1c levels. However, we see that the actual HbA1c is 9.8 but KALP gives a a lower predicted value of 9.0. There are two possible reasons for this. First, Timethoprim is prescribed to treat bacterial infection and the patient would have stopped taking this medication long before her second visit to the ICU, thus removing the negative lab interaction of Timethoprim on HbA1c. Second,

the reported interactions in the AACC, SIDER, MEDI datasets do not include the frequency of occurrence and the severity of the interaction, leading to a less accurate prediction by KALP in this case.