Ontology-aware Prescription Recommendation in Treatment Pathways Using Multi-evidence Healthcare Data

ZIJUN YAO, The University of Kansas, USA
BIN LIU, West Virginia University, USA
FEI WANG, Weill Cornell Medicine, USA
DABY SOW, IBM Research, USA
YING LI, Regeneron, USA

For care of chronic diseases (e.g., depression, diabetes, hypertension), it is critical to identify effective treatment pathways that aim to promptly update the medication following the change of patient state and disease progression. This task is challenging because the optimal treatment pathway for each patient needs to be personalized due to the significant heterogeneity among individuals. Therefore, it is naturally promising to investigate how to use the abundant electronic health records to recommend effective and safe prescriptions. However, prescription recommendation needs to consider multiple aspects of life-critical evidence, such as the information relevance in terms of medical concepts, the health condition in terms of diagnosis history, and the further constraint in terms of side information (e.g., patient demographics and drug side effects). To this end, in this article, we propose a novel prescription recommendation framework named OntoPath to predict the next drug in disease treatment pathways, by building an ontology-aware hierarchical-attention model that integrates multiple medical evidence from domain knowledge guidance, medical history profiling, and side information utilization. Specifically, our method can be characterized from three aspects: (1) by incorporating the longitudinal diagnosis history, we enrich the profiling of patients in terms of comprehensive health conditions, which can largely influence a drug's outcome on individual patients; (2) using the hierarchical disease and drug ontology structures, we are able to model the domain-specific relevance between patients and drugs at multiple levels of granularity and achieve in-depth collaborative filtering; (3) we introduce a pre-training stage to enhance the discriminativeness of network representations, which helps us obtain a premium model initialization to further boost the final recommendation training. We perform extensive experiments on a large-scale depression cohort with over 37,000 patients from a real-world medical claims database. The quantitative and qualitative results demonstrate the effectiveness of OntoPath through the consistent outperformance over state-of-the-art prescription recommendation baselines and the interpretation of model mechanism in case studies.

CCS Concepts: • **Information systems** → **Recommender systems**; • **Applied computing** → **Health informatics;**

Zijun Yao and Ying Li primarily performed the work when they were at Center for Computational Health, IBM Research.
Authors' addresses: Z. Yao (corresponding author), The University of Kansas, 1520 W 15th St, Lawrence, KS 66045, USA; email: zyao@ku.edu; B. Liu, West Virginia University, 83 Beechurst Ave, Morgantown, WV 26505, USA; email: bin.liu1@mail.wvu.edu; F. Wang, Weill Cornell Medicine, 425 E 61st St, New York, NY 10065, USA; email: few2001@med.cornell.edu; D. Sow, IBM Research, 1101 Kitchawan Rd, Yorktown Heights, NY 10598, USA; email: sowdaby@us.ibm.com; Y. Li (corresponding author), Regeneron Pharmaceuticals, Inc., 777 Old Saw Mill River Rd, Tarrytown, NY 10591, USA; email: yl2565@caa.columbia.edu.

1046-8188/2023/04-ART99 $15.00
https://doi.org/10.1145/3579994

Additional Key Words and Phrases: Drug Recommendation, Ontology, Electronic Health Records

ACM Reference format:
Zijun Yao, Bin Liu, Fei Wang, Daby Sow, and Ying Li. 2023. Ontology-aware Prescription Recommendation in Treatment Pathways Using Multi-evidence Healthcare Data. *ACM Trans. Inf. Syst.* 41, 4, Article 99 (April 2023), 29 pages.
https://doi.org/10.1145/3579994

1 INTRODUCTION

Treatment pathway, which refers to a series of prescribing decisions that tailor the medicine for patients over the course of illness [19], plays a critical role in improving the quality and efficiency of patient care. For patients with chronic diseases (e.g., depression, diabetes, hypertension) who generally have a complex pathophysiology during years even decades, it is important to identify a treatment pathway during the journey of patients so that the prescription can be adjusted according to a patient's progression and the disease can be managed in a timely, accurate, and cost-effective manner. However, the treatment pathway can vary from patient to patient due to the wide existence of patient heterogeneity. Taking the treatment pathway of depression patients as an example, Figure 1 shows the transitions between the top-10 most frequent antidepressant drugs given to a 37,000 depression cohort, from the first-line treatment to the third-line treatment, and we can see that there is not a dominant treatment pathway working for everyone. Therefore, the prescription of drugs in a patient's treatment pathway has to to be personalized by considering a variety of real-world evidence [35].

With the ubiquity of the **electronic health records (EHRs)**, creating an intelligent decision support tool for finding treatment pathway has become feasible. Personalized prescription recommendation, which profiles the patient characteristics based on a comprehensive set of medical related evidence such as domain knowledge, medical history, and side information (e.g., patient demographics and drug side effects), has the potential to accurately predict the

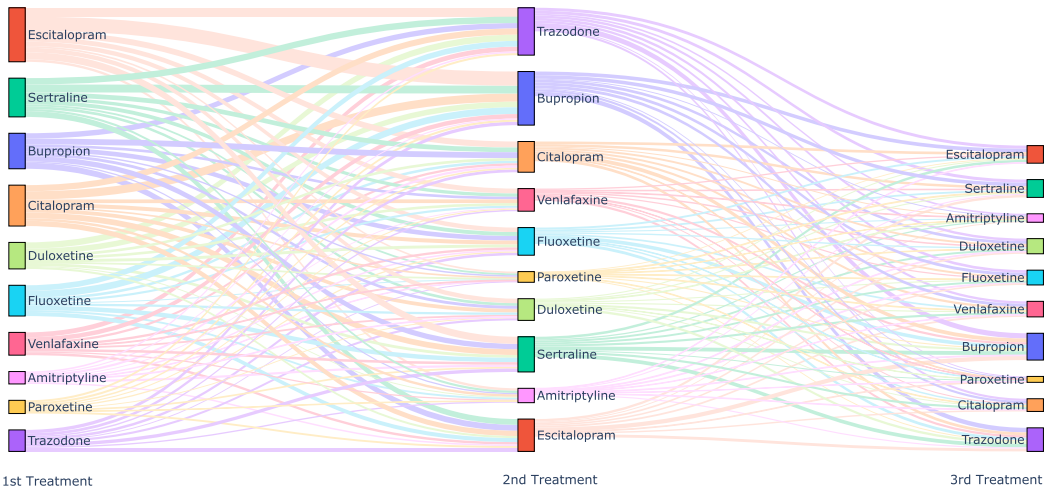


Fig. 1. Treatment pathways of a large-scale depression cohort. The three columns from left to right show the ratio of the top-10 antidepressant prescriptions used for the first-line treatment until the third-line treatment. The link between any two nodes shows the volume of patients switching the prescription from one treatment to another. We can see that since no treatment pathway fits for all the patients, the choice of next treatment has to be personalized.

effectiveness of candidate prescriptions for individual patients. Although literature [37, 50] has shown the promising effectiveness of analyzing longitudinal EHR for prescription prediction, finding prescriptions in treatment pathways still has several unique challenges to tackle. First, rather than predicting the repeated drug observations in doctor visits due to short-term or general symptoms or because of prescription refills, treatment pathways demand a recommendation of the prescriptions that were not given before, and the medication transition is driven by a long-term major disease progression. Second, instead of using a one-size-fits-all model to predict prescriptions for all patients, recommendation for treatment pathways needs to build a personalized model that can profile patients with unique drug adoption, medical history, and demographic information. Last, as a complex decision-making process, we need a comprehensive framework to not only analyze multiple sources of medical evidence but also learn an hierarchical matching function to model the in-depth relevance between patients and drugs. Keeping these challenges in mind, it is appealing to study the three following goals to build the personalized and evidence-based prescription recommendation framework for treatment pathways of chronic disease patients.

First, as the high-stakes decision-making problem, the prediction task of patient-drug effectiveness needs to understand the prescriptions from a domain knowledge perspective. Therefore, the interaction of patient-drug pair can be seen as the interaction between the disease characters from the side of patients and ingredient functions from the side of drugs. To incorporate the domain-specific knowledge into the recommendation model, we utilize the ontological knowledge base regarding human diseases and drugs ingredients to provide the background concepts of EHR observations and the structural relationship among them. For example, built to classify the concepts of disease condition in coarse-to-fine granularity, disease ontology such as ***International Classification of Diseases (ICD)*** is widely used to classify the conditions of different diseases [8]. Similarly, drug ontology such as ***Anatomical Therapeutic Chemical (ATC)*** serves as an important knowledge base to classify the drug ingredients and functions. As shown in Figure 2, when doctors are considering a drug to treat a given medical condition, they would need to know the domain concepts of the disease and the drug and then decide if the two would match. For example, in ICD-9 ontology, the conditions of *Rheumatoid arthritis* is first a connective tissue problem at the top level, then it is described as a joint disorder with inflammatory symptoms at the finer levels of concept. Similarly, for the ingredients of *Dexibuprofen* in ATC ontology, it is first a drug targeting at musculo-skeletal system. Then at finer levels of concept, it has the anti-inflammatory function, with non-steroids and propionic characteristics in ingredients. By incorporating these ontology concepts, we are able to develop the model that mimics how medical practitioners make decisions in matching a patient with a drug using domain-specific knowledge. In our work, the use of ontology bring several advantages in drug recommendation: (1) the hierarchical domain-specific concepts provide a natural way to enrich the representation of patients and drugs; (2) data sparsity issue of rarely appearing information (e.g., rare diseases) can be alleviated by relating them to more generally existing concepts; and (3) multiple levels of knowledge provide the model with a flexibility to choose the most relevant information when considering different patient-drug cases. For example, in Figure 2, considering the *Inflammatory* concept of *Rheumatoid arthritis*, the most relevant information regarding the treatment drug *Dexibuprofen* is the concept of *Anti-inflammatory*.

Second, as a critical step to achieve evidence-based drug recommendation, the profiling of patients should comprehensively capture the longitudinal health conditions from the long-term history in EHR. In most of conventional recommendation use cases (e.g., E-commerce), user profiling can be built by solely looking at the item adoption in history (e.g., what did a user buy in the past define the preference of the user). However, in life-critical prescription decisions, choosing

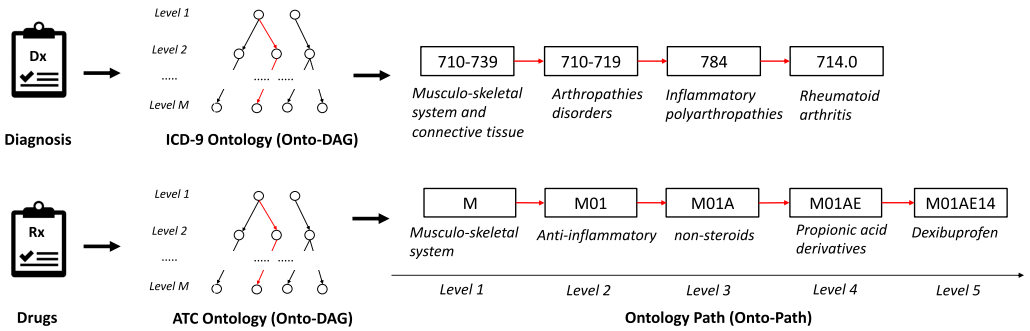


Fig. 2. Illustration of International Classification of Diseases (ICD-9) ontology for diagnosis and Anatomical Therapeutic Chemical (ATC) ontology for drugs. For each diagnosis or drug, we can extract a unique ontology path (Onto-Path) consisting of ontology concepts from the root to the leaf levels of Onto-DAG. The Onto-Path will be used to model the domain-specific interactions of patient-drug pairs.

a drug not only depends on the therapeutic function of the drug specialized for a particular disease but also relies on the personal health state of the patient, like how the disease progresses in the past, and what comorbidities the patient is having in the meantime. Therefore, in prescription recommendation, the medical history with temporal information should be carefully analyzed in the patient profiling. As illustrated in Figure 3 showing the medical history of a sample patient, the longitudinal diagnosis codes indicate the health characteristics of the patient, such as the progression of depression diseases (e.g., Bipolar) and the accompanying comorbidities (e.g., Hypertension). An important job of the recommendation model is to summarize the long-term diagnosis history of each patient and pay attention to the key diagnosis, which can largely decide the prescription outcomes. As a result, by incorporating the longitudinal diagnosis into patient profiling, we are able to create more discriminative patient representations. Therefore, we are more capable of capturing the important relationship between patient states and drugs, and ultimately, it can help the recommendation model to be more trustworthy in practical usage. To this end, how to integrate the temporal diagnosis learning into the patient profiling forms the second goal to be tackled.

Third, from the perspective of optimization, training a recommendation model has typically been treated as a supervised task to predict the observation of a prescription. Although it is an effective way of making the model successful on predicting the recommendation labels, relying on this sole objective may leave the patient or drug representations suboptimal in terms of the discriminativeness in the medical concept space (which is not the primary goal of recommendation classification). This lack of training on representation will ultimately constrain the generalization of recommendation model on the data we have not seen before. From this perspective, if we rethink the recommendation as a matching problem where we would like to see how "close" a patient is to a drug, we can alternatively optimize the recommendation model to represent the relevance of patient-drug pairs as the distance in medical concept embedding space, and use the resulting parameters as a refined initialization for the final recommendation training. With this goal in mind, we propose to adopt an additional self-supervised contrastive learning objective in a pre-training stage for our model, aiming at encoding the medical proximity in the patient and drug representations, for boosting the performance at the final classification training stage for drug recommendation.

Motivated by the aforementioned goals, in this article, we propose a multi-evidence prescription recommendation framework that aims to predict the next drugs in treatment pathways for

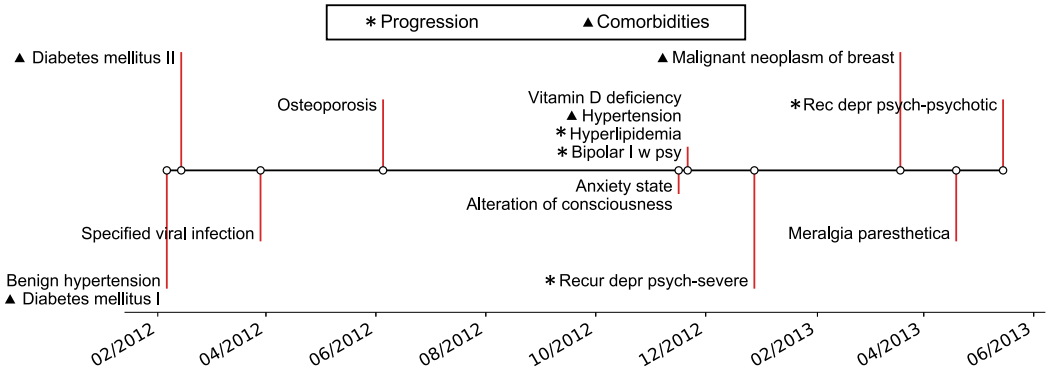


Fig. 3. Longitudinal diagnosis history of an example depression patient. To profile the health condition of patients, the recommendation model needs to summarize the long-term history and focus on the key diagnosis, which can largely influence the outcomes, such as the progression indicators (marked with ∗) and the comorbidities (marked with ▲).

chronic diseases. In this framework, we build a novel collaborative filtering-based recommendation model that simultaneously incorporates evidence from the hierarchical concepts in medical ontology, the longitudinal diagnosis history in EHR, and the important side information of patient demographics and drug side effects. Specifically, we first adopt a demographic-aware self-attention network (i.e., Transformer) to summarize patient health conditions based on longitudinal diagnosis history. To distinguish the key conditions that dominate patient states, we encode each diagnosis by its contextual importance, and decodes the entire diagnosis sequences into the aggregated patient embeddings using their demographic information as query weights. Second, to model a domain-specific in-depth interaction for each patient-drug pair, we utilize a dual-**recurrent neural network (RNN)** network to encode hierarchical ontology concepts and concept structures for patients and drugs, respectively, and use a co-attention with max pooling network to model their level-to-level ontology concept interactions. Third, we adopt the contrastive learning using a self-supervised objective function to pre-train the model for discriminative representation of information. With the pre-trained parameters regarding patients and drugs, we obtain a premium initialization of the network to boost the final optimization of recommendation task. To make the recommendation score more comprehensive, we consider the patients-drugs prediction in two parts—the therapeutic relevance through drug function, and the safe relevance through drug side effects. Eventually, all the sub-components mentioned above are integrated as a whole model that is trained in an end-to-end manner.

In evaluation, we validate the proposed OntoPath prescription recommendation model through extensive experiments on a large-scale depression patient cohort extracted from a real-world medical claims database (i.e., MarketScan). The results show that OntoPath consistently outperforms baseline models by a significant margin and demonstrate the effectiveness of each model subcomponent. Furthermore, we conduct a qualitative study to show that the attention mechanism in our model is able to recover clinically meaningful insights, which enable the explainability character of the proposed framework.

2 PROBLEM FORMULATION

We first introduce two key concepts and then formulate our prescription recommendation problem.

Table 1. Mathematical Notations

Symbol	Description
u, i, d	Index for patients, drugs, or diagnosis
$c \in \mathcal{G}_{icd}$, $a \in \mathcal{G}_{atc}$	Concept node c in ICD-9 Onto-DAG, or concept node a in ATC Onto-DAG
M_{icd}, M_{atc}	Depth of ICD-9, or ATC Onto-DAGs
$\mathcal{P}^{(d)} = \{c_1, \ldots, c_{M_{icd}}\}$	Onto-Path of diagnosis d consisting of ICD-9 concept nodes
$\mathcal{P}^{(i)} = \{a_1, \ldots, a_{M_{atc}}\}$	Onto-Path of drug i consisting of ATC concept nodes
$o^{(u)}$, $e^{(i)}$	Demographic vector of patient u, or side effect vector of drug i
$\boldsymbol{V}_z^{(u)} = \{\boldsymbol{v}_{z_1}^{(u)}, \ldots, \boldsymbol{v}_{z_{M_{icd}}}^{(u)}\}$	Hierarchical embedding of patient u consisting of M_{icd} vectors
$\boldsymbol{V}_a^{(i)} = \{\boldsymbol{v}_{a_1}^{(i)}, \ldots \boldsymbol{v}_{a_{M_{atc}}}^{(i)}\}$	Hierarchical embedding of drug i consisting of M_{atc} vectors
$\boldsymbol{s}^{(u)}$, $\boldsymbol{s}^{(i)}$	Aggregated embedding of patient or drug after hierarchical interaction

Definition 1 (Onto-DAG). An ***Ontology Directed Acyclic Graph (Onto-DAG)*** $\mathcal{G}$ is a hierarchical structure to organize medical concepts with different granularity. In an Onto-DAG, each leaf node usually represents a specific medical code such as a disease diagnosis or a drug ingredient, each internal node usually represents a classification concept to categorize descendant nodes, and each directed edge indicates a patient-to-child relationship where parent nodes provide more general classification and child nodes provide more specific classification.

Definition 2 (Onto-Path). In an Onto-DAG $\mathcal{G}$, every diagnosis or drug can be represented by a unique ***Ontology Path (Onto-Path)*** $\mathcal{P}$, which is a sequence of ontology concepts that traverses through every level of $\mathcal{G}$ from root to leaf, starting from the most general classification concept and ending at the most specific code as shown in Figure 2. For instance, a drug i's Onto-Path can be represented as $\mathcal{P}^{(i)} = a_1^{(i)} \rightarrow a_2^{(i)} \rightarrow \cdots \rightarrow a_{M_{atc}}^{(i)}$, where $a_m^{(i)}$ means the ATC concept of drug i on level m in ATC Onto-DAG $\mathcal{G}_{atc}$, and M_{atc} means the depth of $\mathcal{G}_{atc}$.

Problem Definition. Suppose we have a set of patients $u \in \mathcal{U}$, a set of drugs $i \in \mathcal{I}$, and the observed prescriptions of patient-drug pairs $y^{(u,i)}$. Meanwhile, we have a set of diagnosis $d \in \mathcal{D}$, and each patient u's medical history can be represented as a sequence of unique diagnosis codes $\{d_1^{(u)}, \ldots, d_T^{(u)}\}$, where diagnosis are ordered by their earliest appearance time t for patient u. In addition, we have two Onto-DAGs: $\mathcal{G}_{atc}$ of ATC ontology shows the therapeutic concepts of active ingredients of drugs, and $\mathcal{G}_{icd}$ of ICD-9 ontology shows the diagnostic concepts of diseases. Given the Onto-DAGs, we can extract a unique Onto-Path $\mathcal{P}^{(i)} = \{a_1^{(i)}, a_2^{(i)}, \ldots, a_{M_{atc}}^{(i)}\}$ for each drug i and a unique Onto-Path $\mathcal{P}^{(d)} = \{c_1^{(d)}, c_2^{(d)}, \ldots, c_{M_{icd}}^{(d)}\}$ for each diagnosis d, showing the hierarchical concepts from general to specific categorization with a total of M_{atc} or M_{icd} levels. In addition, we have the vector of side information for demographics of each patient $o^{(u)}$ such as age/sex, geographic region, and employment class/status, and for side effects of each drug $e^{(i)}$ covering all the possible drug adverse reactions such as allergy, high-temperature, and cardiac arrhythmias.
Task: Given (1) every patient's prescription history and medical history from the first visit to the latest visit at time T; (2) the Onto-Paths $\mathcal{P}^{(d)}$ and $\mathcal{P}^{(i)}$ for each diagnosis and drug; and (3) the demographics $o^{(u)}$ and side effects $e^{(i)}$ features for each patient and drug, the goal is to recommend the new prescription (i.e., the drugs have not been adopted in the patient's history) for the next visit at $T+1$ of individual patients. Table 1 summarizes the important notations in our framework.

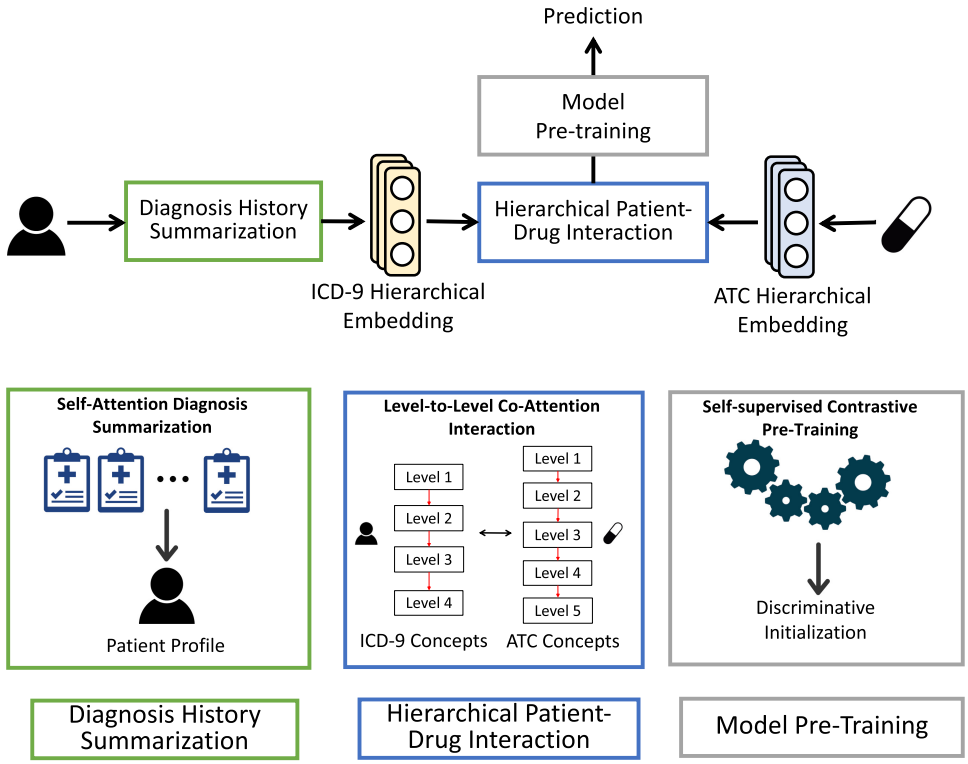


Fig. 4. Framework overview. The model architecture consists of three major components. Green: Patients are profiled by summarizing personal diagnosis history. Blue: Patients and drugs are interacted based on hierarchical ontology embeddings. Grey: Model parameters are pre-trained for discriminative initialization.

3 METHODOLOGY

3.1 Framework Overview

Figure 4 shows the overview of OntoPath recommendation framework. Generally, the architecture consists of three components: (1) a module for diagnosis history summarization will analyze the sequence of diagnosis and aggregate them as unified patient embeddings. Specifically, by viewing each diagnosis as hierarchical ontology concepts in a ICD-9 Onto-Path, we perform the history summarization separately at different levels, from the root-level concepts to the leaf-level concepts. Eventually, for each patient, we will obtain M_{icd} (e.g., 4) hierarchical embeddings showing the general-to-specific profile of patients; (2) a hierarchical patient-drug interaction module will accept the ontology-driven hierarchical embeddings of patients and drugs, and then will formulate their level-to-level interactions to model the prescribing decisions comprehensively. Specifically, we use a dual-RNN model to encode the ontology structures and use a co-attention mechanism to facilitate level-to-level patient-drug matching; and (3) a pre-training module will be conducted to learn a self-supervised contrastive loss ahead of the final recommendation training. As a result, we will obtain a discriminative initialization of model parameters such as patient/drug representations to boost the final training of prescription prediction. Last, we compute the recommendation score supervised by the label of prescriptions in training set, where positive prescriptions are marked 1 and negative prescriptions are marked 0.

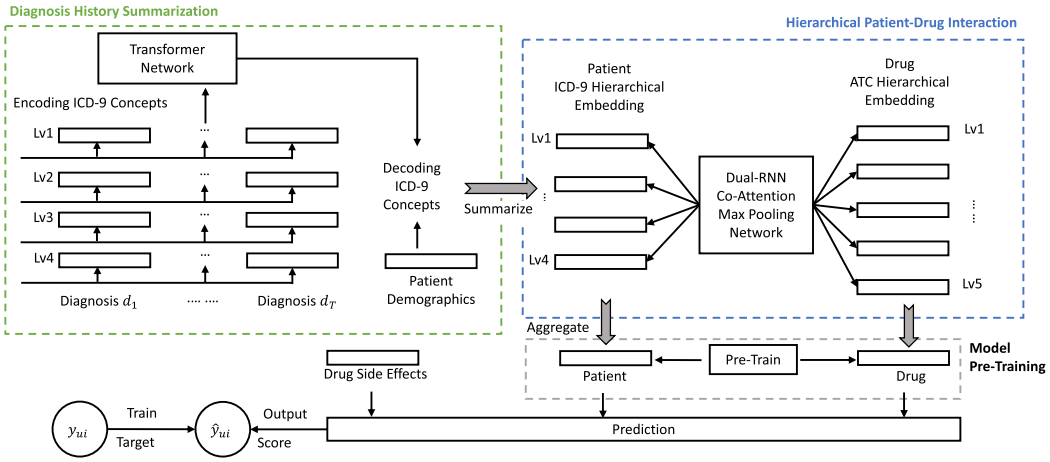


Fig. 5. Details of OntoPath, an end-to-end prescription recommendation model. Three major model components are highlighted by dashed box in different colors. Green box shows the diagnosis history summerization module for patient profiling. Blue box indicates hierarchical patient-drug interaction module to facilitate level-of-level co-attentions. Grey box illustrate the model pre-training module to refine the representation initialization for the final recommender training.

In the following sections, we will present the details of the aforementioned model components following the zoom-in illustration in Figure 5. First, we explain how to summarize the long diagnosis history hierarchically, with the incorporation of ICD-9 Onto-Paths using the self-attention model – a Transformer network decoded by demographics. Second, we elaborate how to utilize both the ICD-9 hierarchical embedding of patients and the ATC hierarchical embedding of drugs to conduct level-to-level interactions using a dual-RNN co-attention network. Third, we introduce how to optimize the model with a contrastive learning loss in a pre-training stage to provide a premium initialization for boosting the learning of recommendation labels. Finally, we demonstrate the final prediction layer, which learns the prescribing decision for each patient-drug pair by considering two concerns: the therapeutic effectiveness (using the ingredient knowledge of drugs) and the therapeutic safety (using the side-effect knowledge of drugs).

3.2 Profiling Patients by Summarizing Diagnosis History

In this section, we profile patients by summarizing the sequence of diagnosis in history. Specifically, we aim to learn the patient hierarchical embeddings, where each level aggregates the ICD-9 concepts of all diagnosis with a certain ontology granularity. The idea of learning diagnosis and aggregating them as patient embedding is motivated by (1) diagnosis is an indispensable evidence characterizing patient conditions before any treatment; and (2) diagnosis history provide us with an natural way to introduce domain-specific concepts (ICD-9) into patient representation, where each patient can be summarized in conditions at different levels of granularity using hierarchical embeddings. To this end, as illustrated in green dashed box in Figure 5, we adopt an Transformer network [40] to learn the sequence of diagnosis by processing their ontology concepts simultaneously at all granularity levels (encoding). Then, we use demographic information as the query weights to aggregate the diagnosis sequence (decoding) to profile patient conditions with personalized awareness.

3.2.1 Encoding Diagnosis with Self-attention. Given a patient u's diagnosis history $\{d_1^{(u)}, \ldots, d_T^{(u)}\}$ where each diagnosis d has an Onto-Path $\mathcal{P}^{(d)} = \{c_1^{(d)}, \ldots, c_{M_{icd}}^{(d)}\}$ with a

length M_{icd}, we can extract M_{icd} (e.g., 4) ontology concept sequences from the diagnosis history where each one focuses on a particular level of ICD-9 ontology. For example, given a sequence of diagnosis in terms of ICD-9 codes [*296.90*, *784.0*, *031.1*], we can represent this sequence as four sequences of ICD-9 concepts at different levels. For example, the first level will be [*290-319*, *780-799*, *001-139*], classifying the organ or system of the diagnosis; the second level [*295-299*, *780-789*, *030-041*] and the third level [*296*, *784*, *031*] will show more specific diagnosis categorization; the last level of ICD-9 concepts are the diagnosis codes themselves.[1] Formally, we can express the hierarchical representation of diagnosis sequence as follows:

$$\begin{bmatrix} \mathcal{P}^{(d_1)} & \mathcal{P}^{(d_2)} & \cdots & \mathcal{P}^{(d_T)} \end{bmatrix} = \begin{bmatrix} c_1^{(d_1)} & c_1^{(d_2)} & \cdots & c_1^{(d_T)} \\ c_2^{(d_1)} & c_2^{(d_2)} & \cdots & c_2^{(d_T)} \\ c_3^{(d_1)} & c_3^{(d_2)} & \cdots & c_3^{(d_T)} \\ c_4^{(d_1)} & c_4^{(d_2)} & \cdots & c_4^{(d_T)} \end{bmatrix}, \tag{1}$$

where $\mathcal{P}^{(d)}$ means the Onto-Path of the diagnosis d from 1 to T in the patient's medical history. For example, ICD Onto-Path of the first diagnosis d_1 can be represented by four ICD concepts $\mathcal{P}^{(d_1)} = \begin{bmatrix} c_1^{(d_1)} & c_2^{(d_1)} & \cdots & c_4^{(d_1)} \end{bmatrix}^\top$, therefore, a sequence of diagnosis can be represented hierarchically by four rows of concepts in Equation (1).

Then, for each row of the hierarchical representation, we will process a sequence consisting of T ICD-9 concepts $c^{(d_1)}, c^{(d_2)}, \ldots, c^{(d_T)}$ (we omit the subscript for levels here). By having an embedding for each ICD-9 concept, we denote the sequence of ICD-9 concepts as matrix $\boldsymbol{C'} \in \mathbb{R}^{T \times K}$ where each row $\boldsymbol{v}'_{c_t}$ hosts the K-dimensional embedding of the ICD-9 concept at time t. Since Transformer network does not have a recurrent setting, first we need to add the position information $\boldsymbol{p}_t$ to $\boldsymbol{v}'_{c_t}$:

$$\begin{aligned} p_{t,2i} &= \sin(t/10,000^{2i/K}), \\ p_{t,2i+1} &= \cos(t/10,000^{2i/K}), \end{aligned} \tag{2}$$

where $t \in [0, T)$ indicates the time step in diagnosis history and i indicates the dimension index in K. By combining the ICD-9 concept embedding and sequential position embedding, we obtain the input embeddings $\boldsymbol{C} \in \mathbb{R}^{T \times K}$ where each row $\boldsymbol{v}_{c_t} = \boldsymbol{v}'_{c_t} \times \sqrt{K} + \boldsymbol{p}_t$.

Next, given the input embedding of ICD-9 concept sequence $\boldsymbol{C}$, we aim to encode it with the multi-head self-attention function of Transformer network. Suppose we have N heads attention, every head of attention will independently encodes each ICD-9 concept into a subspace based on the attentions given by the rest of ICD-9 concepts in the same sequence. Specifically, for the nth single-head, we encode $\boldsymbol{C}$ into three embeddings: query $\boldsymbol{Q}_n^{(C)} = \boldsymbol{C}\boldsymbol{W}_n^{(qry)}$, key $\boldsymbol{K}_n^{(C)} = \boldsymbol{C}\boldsymbol{W}_n^{(key)}$ and value $\boldsymbol{V}_n^{(C)} = \boldsymbol{C}\boldsymbol{W}_n^{(val)}$, where $\boldsymbol{W}_n^{(qry)}, \boldsymbol{W}_n^{(key)}, \boldsymbol{W}_n^{(val)} \in \mathbb{R}^{K \times K/N}$ are the trainable parameters for the nth head of attention. In the end, we obtain $\boldsymbol{A}_n^{(C)} \in \mathbb{R}^{T \times K/N}$ as the encoding results of $\boldsymbol{C}$ with the attention generated by a single head function:

$$\boldsymbol{A}_n^{(C)} = \text{softmax}\left(\frac{\boldsymbol{Q}_n^{(C)} \boldsymbol{K}_n^{(C)\top}}{\sqrt{d}}\right) \boldsymbol{V}_n^{(C)}, \tag{3}$$

where $d = K/N$ means the dimension of each single head embedding of ICD-9 concepts.

With the concatenation of the embedding from all attention heads and a following fully connected layer (**multilayer perceptron (MLP)**), we obtain the overall encoded results $\boldsymbol{A}^{(C)}$, which

[1]Since most diagnosis codes in ICD-9 ontology are at the depth of 4 or 5, we uniformly have $M_{icd} = 4$ for all Onto-Paths of diagnosis code consisting of the three topmost categorical concepts and one leaf concept.

has the same dimension with the input $\boldsymbol{C}$ but each ICD-9 concept in the sequence has been evaluated based on the attentions from the context concepts:

$$\boldsymbol{A}^{(C)} = \text{MLP}\left(\text{concat}\left(\boldsymbol{A}_1^{(C)}, \ldots, \boldsymbol{A}_N^{(C)}\right)\right). \tag{4}$$

3.2.2 Decoding Diagnosis with Demographic-attention. Given $\boldsymbol{A}^{(C)}$ the self-attention encoding of every ICD-9 concept in sequence input $\boldsymbol{C}$, we will summarize the entire sequence into a single embedding reflecting the patient health conditions. We use the decoding function of Transformer to aggregate the ICD-9 concepts in $\boldsymbol{A}^{(C)}$. Particularly, we introduce the patient demographic features to generate personalized attentions for aggregation. The idea is that, by different age, sex, locations, socioeconomic status, and lifestyles, a particular patient should have different strategy to assess health conditions (e.g., diabetes can be more dangerous for seniors), which would largely influence the treatment to choose (e.g., treatment with less aggressive side effects). In detail, we map the patient u' demographic vector $\boldsymbol{o}^{(u)} \in \mathbb{R}^{1\times K}$ to the query $\boldsymbol{q}_n^{(o)} = \boldsymbol{o}^{(u)}\boldsymbol{W}_n^{'(qry)}$, and map the formerly encoded diagnosis concepts in the sequence $\boldsymbol{A}^{(C)}$ to the key and the value: $\boldsymbol{K}_n^{(A)} = \boldsymbol{A}^{(C)}\boldsymbol{W}_n^{'(key)}$ and $\boldsymbol{V}_n^{(A)} = \boldsymbol{A}^{(C)}\boldsymbol{W}_n^{'(val)}$, where $\boldsymbol{W}_n^{'(qry)}, \boldsymbol{W}_n^{'(key)}, \boldsymbol{W}_n^{'(val)} \in \mathbb{R}^{K\times K/N}$ stand for the parameters of Transform decoder. Then, we get the aggregated embedding $\boldsymbol{v}_{z,n}^{(u)} \in \mathbb{R}^{1\times K/N}$ of the patient u's health conditions summarized by the nth head of attentions:

$$\boldsymbol{v}_{z,n}^{(u)} = \text{softmax}\left(\frac{\boldsymbol{q}_n^{(o)}\boldsymbol{K}_n^{(A)\top}}{\sqrt{d}}\right)\boldsymbol{V}_n^{(A)}. \tag{5}$$

By concatenating the resulting embeddings from all the N attention heads and process it through a MLP layer, we get the patient embedding $\boldsymbol{v}_z^{(u)} \in \mathbb{R}^{1\times K}$ summarizing a sequence of ICD-9 concepts at a particular ontology level:

$$\boldsymbol{v}_z^{(u)} = \text{MLP}\left(\text{concat}\left(\boldsymbol{v}_{z,1}^{(u)}, \ldots, \boldsymbol{v}_{z,N}^{(u)}\right)\right). \tag{6}$$

For each patient u's diagnosis history, we have four sequences of concepts to process because of the length $M_{icd} = 4$ of each ICD-9 Onto-Path. Eventually, by finishing the sequence of ICD-9 concepts at all levels, we obtain the hierarchical embedding of patient u as $\boldsymbol{V}_z^{(u)} \in \mathbb{R}^{4\times K}$:

$$\boldsymbol{V}_z^{(u)} = \begin{bmatrix} \boldsymbol{v}_{z_1}^{(u)} \\ \ldots \\ \boldsymbol{v}_{z_4}^{(u)} \end{bmatrix}. \tag{7}$$

3.3 Modeling Patient-Drug Interactions using Hierarchical Embeddings

Once we have obtained the hierarchical embedding of patients, we are ready to formulate the interactions between patients and drugs. First, with the aforementioned ATC Onto-Path of drugs $\mathcal{P}^{(i)}$, we can hierarchically represent each drug as M_{atc} (e.g., 5) levels of ATC concepts $\mathcal{P}^{(i)} = \begin{bmatrix} a_1^{(i)} & a_2^{(i)} & \ldots & a_5^{(i)} \end{bmatrix}^\top$. Given each ATC concept an embedding, we formally represent the hierarchical embedding of drug[2] i as $\boldsymbol{V}_a^{(i)} \in \mathbb{R}^{5\times K}$:

$$\boldsymbol{V}_a^{(i)} = \begin{bmatrix} \boldsymbol{v}_{a_1}^{(i)} \\ \ldots \\ \boldsymbol{v}_{a_5}^{(i)} \end{bmatrix}. \tag{8}$$

[2]All drugs in ATC ontology have exactly five levels of ingredient and therapeutic concepts.

We aim to find out how to match a patient $\boldsymbol{V}_z^{(u)}$ with a drug $\boldsymbol{V}_a^{(i)}$ so that the model can best explain the observation of prescriptions. Therefore, given the hierarchical embeddings, we formulate the interaction between the two sides through a co-attention mechanism across different levels of ontology concepts. In this way, the patient-drug matching can be decomposed into a series of hierarchical ontology concept matching, which can better address the final recommendation decisions. For example, when we are considering a drug for a certain purpose of treatment, we start from the most general ATC classification concepts (e.g., *anti-infectives*), then we examine the more specific concepts (e.g., *anti-bacterials*) until we reach the exact active ingredient of the drug (e.g., *Penicillin*). Similarly, when we are considering a patient with a key condition, we check through the general-to-specific ICD-9 classification concepts to identify the problems, like *neoplasms → eoplasms of lymphatic tissue → acute lymphoid leukemia*. With this motivation, we propose to achieve the function in two steps: (1) encoding hierarchical embedding from general to specific levels to preserve the ontology structure; and (2) modeling interactions between patient and drug with a level-to-level co-attention mechanism. This component is illustrated by the blue dashed box in Figure 5.

3.3.1 Encoding Ontology Structure with Recurrent Networks. To encode the embedding of concepts in $\boldsymbol{V}_z^{(u)}$, $\boldsymbol{V}_a^{(i)}$ with ontology structure information (e.g., parent to child relationship), we adopt dual-RNN for patients and drugs to encode the ICD-9 and ATC concepts separately. Specifically, for each patient-drug pair (u, i), we have the hierarchical embeddings of patient $\boldsymbol{V}_z^{(u)} = \{\boldsymbol{v}_{z_1}^{(u)}, \ldots, \boldsymbol{v}_{z_{M_{icd}}}^{(u)}\}$ and drug $\boldsymbol{V}_a^{(i)} = \{\boldsymbol{v}_{a_1}^{(i)}, \ldots \boldsymbol{v}_{a_{M_{atc}}}^{(i)}\}$, where $M_{icd} = 4$, $M_{atc} = 5$, and $\boldsymbol{v}_{z_m}^{(u)}, \boldsymbol{v}_{a_m}^{(i)} \in \mathbb{R}^{1\times K}$. We recurrently input the hierarchical embeddings in a top-to-bottom order to a RNN, so that each level of ontology concepts will include the information of their ancestors.

We obtain the embedding of patient and drug information at each level m from two separate RNN:

$$\begin{aligned} \boldsymbol{h}_{z_m}^{(u)} &= \tanh(\boldsymbol{h}_{z_{m-1}}^{(u)} \boldsymbol{W}_{hh}^{(icd)} + \boldsymbol{v}_{z_m}^{(u)} \boldsymbol{W}_{xh}^{(icd)} + \boldsymbol{b}^{(icd)}), \\ \boldsymbol{h}_{a_m}^{(i)} &= \tanh(\boldsymbol{h}_{a_{m-1}}^{(i)} \boldsymbol{W}_{hh}^{(atc)} + \boldsymbol{v}_{a_m}^{(i)} \boldsymbol{W}_{xh}^{(atc)} + \boldsymbol{b}^{(atc)}), \end{aligned} \tag{9}$$

where $\boldsymbol{h}_{z_m}^{(u)}, \boldsymbol{h}_{a_m}^{(i)} \in \mathbb{R}^K$ are the ontology structure encoded embeddings of patient u and drug i at mth level. $\boldsymbol{W}_{hh}^{(icd)}, \boldsymbol{W}_{xh}^{(icd)}, \boldsymbol{W}_{hh}^{(atc)}, \boldsymbol{W}_{xh}^{(atc)} \in \mathbb{R}^{K\times K}$ and $\boldsymbol{b}^{(icd)}, \boldsymbol{b}^{(atc)} \in \mathbb{R}^{1\times K}$ are the network parameters for both recurrent networks. It is worth noting that we use the vanilla RNN model to simplify the illustration, more complex version of RNN like **Gated Recurrent Unit (GRU)** or **Long Short-term Memory (LSTM)** can be adopted as well for better preserving the ontology structures in this process. In this work, we used GRU in our experiments.

3.3.2 Inferring Level-to-Level Co-attention between Patients and Drugs. Through the dual-RNN encoder, we obtain the final hierarchical embedding of patient and drug in terms of ontology concepts and structures, next we start to model the mutual level-to-level interaction between them. Having the health condition or the treatment function explained by multiple levels of patient or drug embeddings, different levels should be emphasized when different patient-drug pairs are considered. Therefore, we aim to make the level-wise attention, which is dynamically adjusted to different patient-drug examples, so that we can flexibly model the decisions of the final prescription. Motivated by this intuition, we develop a co-attention network based on Reference [36] to organize the level-to-level interaction between the patient and drug hierarchical embeddings.

Given the hierarchical embeddings output from Equation (9) in dual-RNN encoder: patient $\boldsymbol{H}_z^{(u)} \in \mathbb{R}^{M_{icd}\times K}$ and drug $\boldsymbol{H}_a^{(i)} \in \mathbb{R}^{M_{atc}\times K}$, where each row in $\boldsymbol{H}_z^{(u)}$ or $\boldsymbol{H}_a^{(i)}$ corresponds to the mth level ontology concepts $\boldsymbol{h}_{z_m}^{(u)}$ or $\boldsymbol{h}_{a_m}^{(i)}$. We compute an affinity matrix $\boldsymbol{G}^{(u,i)} \in \mathbb{R}^{M_{icd}\times M_{atc}}$ to

accommodate the level-to-level concept relevance:

$$\boldsymbol{G}^{(u,i)} = \tanh(\boldsymbol{H}_z^{(u)} \Sigma \boldsymbol{H}_a^{(i)\top}), \tag{10}$$

where $\boldsymbol{G}^{(u,i)}$ provide the co-attention alignment scores to the pairs of concepts, which belong to the patient u and the drug i. For example, $G_{3,4}^{(u,i)}$ shows the alignment score between the thrid-level concept of patient and the fourth-level concept of drug. $\Sigma \in \mathbb{R}^{K \times K}$ is a matrix of parameters, for coordinating the interactions across embedding dimensions. In this way, given $\boldsymbol{G}$, we let either patient u or drug i to mutually decide the attention coefficients for each other.

Next, we apply a max pooling function by taking the maximum value over columns (all levels of drug) and rows (all levels of patient) of $\boldsymbol{G}^{(u,i)}$ to obtain the overall alignment scores:

$$\boldsymbol{g}^{(u)} = \max_{col}(\boldsymbol{G}^{(u,i)}) \quad \text{and} \quad \boldsymbol{g}^{(i)} = \max_{row}(\boldsymbol{G}^{(u,i)}), \tag{11}$$

where $\boldsymbol{g^{(u)}} \in \mathbb{R}^{M_{icd}}, \boldsymbol{g^{(i)}} \in \mathbb{R}^{M_{atc}}$ are the final alignment scores of each level for patient and drug, respectively. During this process, since $\boldsymbol{G}$ depends on the specific concepts appearing in the patient or the drug hierarchical embeddings, the final attention weights dynamically change according to the actual patient-drug examples.

Finally, we calculate the attention weights $\boldsymbol{\alpha}^{(u)}$ and $\boldsymbol{\alpha}^{(i)}$ of each level m in hierarchical embeddings using softmax function:

$$\begin{aligned} \alpha_m^{(u)} &= \frac{\exp(g_m^{(u)})}{\sum_{1 \le l \le M_{icd}} \exp(g_l^{(u)})}, \\ \alpha_m^{(i)} &= \frac{\exp(g_m^{(i)})}{\sum_{1 \le l \le M_{atc}} \exp(g_l^{(i)})}. \end{aligned} \tag{12}$$

With the level-wise attention weights on the input hierarchical embeddings $\boldsymbol{H}_z^{(u)}$, $\boldsymbol{H}_a^{(u)}$, we obtain the aggregation of hierarchical embeddings $\boldsymbol{s}^{(u)}, \boldsymbol{s}^{(i)} \in \mathbb{R}^{1 \times K}$ of patients and drugs:

$$\boldsymbol{s}^{(u)} = \boldsymbol{\alpha}^{(u)\top} \boldsymbol{H}_z^{(u)} \quad \text{and} \quad \boldsymbol{s}^{(i)} = \boldsymbol{\alpha}^{(i)\top} \boldsymbol{H}_a^{(i)}. \tag{13}$$

3.4 Pre-training Model with Contrastive Learning

As shown by the grey dashed box in Figure 5, the level aggregated embedding $\boldsymbol{s}^{(u)}$, $\boldsymbol{s}^{(i)}$ of patient u and drug i obtained from in Equation (13) are the inputs for the prediction layer of recommendation. Intuitively, we can simply use a supervised objective function such as a classification or regression loss to approach the label of patient-drug observations and optimize the network parameters. However, the direct optimization for the supervision label sometimes overlooks the discriminativeness of representations such as $\boldsymbol{s}^{(u)}$ and $\boldsymbol{s}^{(i)}$. For example, given a patient-drug pair with a strong relevance that they should be recommended, the representation of $\boldsymbol{s}^{(u)}$ and $\boldsymbol{s}^{(i)}$ should also express a high intimacy in the semantic embedding space to reflect a good matching score for this pair. To this end, we adopt a loss of contrastive learning [5] to pre-train the discriminativeness of $\boldsymbol{s}^{(u)}$ and $\boldsymbol{s}^{(i)}$, so that we can obtain a premium initialization of model parameters prepared for the final recommendation label learning. Specifically, we use the contrastive loss to distinguish if a sequence of diagnosis history and a drug prescription belong to the same patient's records. In our work, these two representations for contrasting are exactly $\boldsymbol{s}^{(u)}$ and $\boldsymbol{s}^{(i)}$.

Formally, by randomly sampling a minibatch of N patient-drug pairs $\{(\boldsymbol{s}_n^{(u)}, \boldsymbol{s}_n^{(i)})\}_{n=1}^{N}$, we define the contrastive loss as

$$\mathcal{L}_{\text{pre-train}} = -\frac{1}{N} \sum_{n=1}^{N} \log \frac{\exp(\text{sim}(\boldsymbol{s}_n^{(u)}, \boldsymbol{s}_n^{(i)})/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\boldsymbol{s}_n^{(u)}, \boldsymbol{s}_k^{(i)})/\tau)}, \tag{14}$$

ALGORITHM 1: The algorithm of OntoPath model.

Input: Patient u, drug i, diagnosis history $\{d_1^{(u)}, \ldots, d_T^{(u)}\}$, ICD-9 Onto-Path $\mathcal{P}^{(d)}$ of diagnosis d, ATC Onto-Path of drug i, patient demographics $o^{(u)}$ and drug side effects $i^{(i)}$.
Output: Recommendation score $\hat{y}^{(u,i)}$.

```
Initialize all parameters
while not reaching convergence criteria do
    for each patient-drug instance y^(u,i) in a mini-batch do
        # Diagnosis history summarization
        Represent patient's diagnosis history as 4 hierarchical ICD-9 concept sequences by Equation (1)
        for ICD-9 concept sequences at each ontology level do
            Encode ICD-9 concepts in the sequence through self-attention by Equations (3), (4)
            Decode the sequence of ICD-9 concepts with demographics by Equations (5), (6)
        end for
        # Hierarchical patient-drug interaction
        Encode ontology structure of patient and drug with dual-RNN by Equation (9)
        Level-to-level max-pooling co-attention between patient and drug by Equations (10), (11), (12)
        Obtain aggregation of hierarchical embedding for patient and drug by Equation (13)
    end for
    # Model optimization
    if model pre-training then
        Calculate contrastive loss and update parameters by Equation (14)
    else
        Calculate recommendation score in terms of effectiveness and safety by Equation (15)
        Calculate cross-entropy loss and update parameters by Equation (16)
    end if
end while
```

where n denotes the nth patient-drug pair in observation and N is the total number of pairs in the minibatch. For the particular nth patient-drug observation, we have 1 positive sample $(\boldsymbol{s}_n^{(u)}, \boldsymbol{s}_n^{(i)})$, and $N-1$ negative samples $(\boldsymbol{s}_n^{(u)}, \boldsymbol{s}_k^{(i)})$ where $k \neq n$, because $\boldsymbol{s}_k^{(i)}$ comes from other patients. sim() means the similarity function such as cosine similarity or dot product, τ is the hyperparameter for softmax temperature.

3.5 Learning Final Patient-Drug Recommendation

After the pre-train stage, we prepare the inputs for the final recommendation label prediction. We have the pre-trained embeddings of patients $\boldsymbol{s}^{(u)}$ versus the pre-trained embedding of drugs $\boldsymbol{s}^{(i)}$ and an additional vector $e^{(i)}$—the side-effect features of drugs. The motivation is that we attempt to model two perspectives of relevance between the patients and the drugs: (1) the effectiveness by interacting patients with the therapeutic characteristics of drugs ($\boldsymbol{s}^{(i)}$ obtained from ATC ontology concepts); and (2) the safety by interacting patients with the adverse reaction of drugs ($e^{(i)}$ obtained from side effects).

For each observed patient-drug pairs (labeled $y = 1$), we sample (e.g., $N_{ns} = 5$) negative pairs (labeled $y = 0$) by randomly replacing the drug with others. We use element-wise product to obtain the embedding of effectiveness and safety relevance, and have it processed through a classification

output layer to generate the classification probability:

$$\hat{y}^{(u,i)} = \sigma_{\text{sigmoid}}(\boldsymbol{w}^{(out)\top}\text{concat}((\boldsymbol{s}^{(u)} \odot \boldsymbol{s}^{(i)}), (\boldsymbol{s}^{(u)} \odot e^{(i)}))), \tag{15}$$

where $\odot$ means the element-wise product of embeddings, $\sigma_{\text{sigmoid}}(x) = 1/(1 + e^{-x})$ is the sigmoid function, and $\boldsymbol{w}^{(out)}$ are the parameters for output layer.

Finally, we adopt the cross-entropy loss for optimizing over all the positive and negative patient-drug pairs:

$$\mathcal{L}_{\text{rec}} = -\sum_{(u,i)} y^{(u,i)} \log \hat{y}^{(u,i)} + (1 - y^{(u,i)}) \log(1 - \hat{y}^{(u,i)}), \tag{16}$$

where $y^{(u,i)}$ is the positive or negative prescription label 1 or 0, and $\hat{y}^{(u,i)}$ is the predicted probability of positive class.

Algorithm 1 shows the training process of recommendation model. Once the training is finished, for each new patient-drug pair, we are able to generate its recommendation score $\hat{y}^{(u,i)}$. Therefore, in the testing stage, given a patient and a list of candidate drugs for consideration, we rank the drugs based on their recommendation scores in descending order. The top-k drug candidates will be the recommendations for the patient.

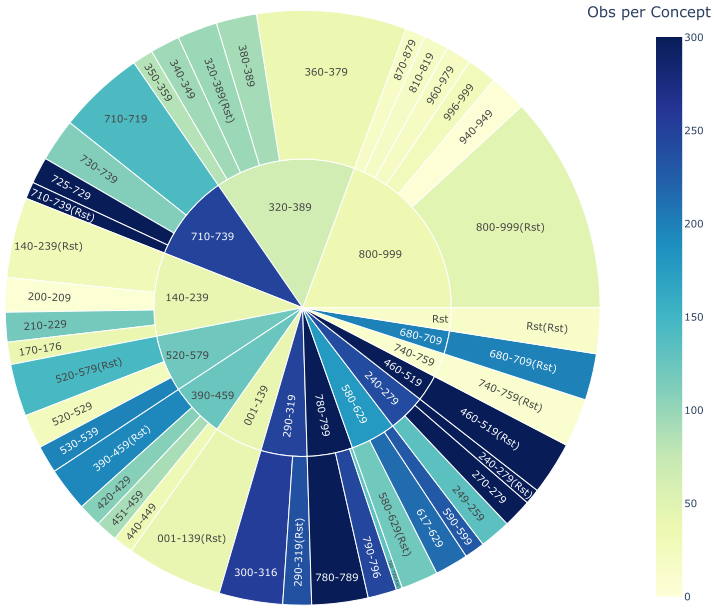
4 EXPERIMENTS

In this section, we empirically evaluate the performance of the proposed OntoPath prescription recommendation framework on a real-world medical claims dataset.

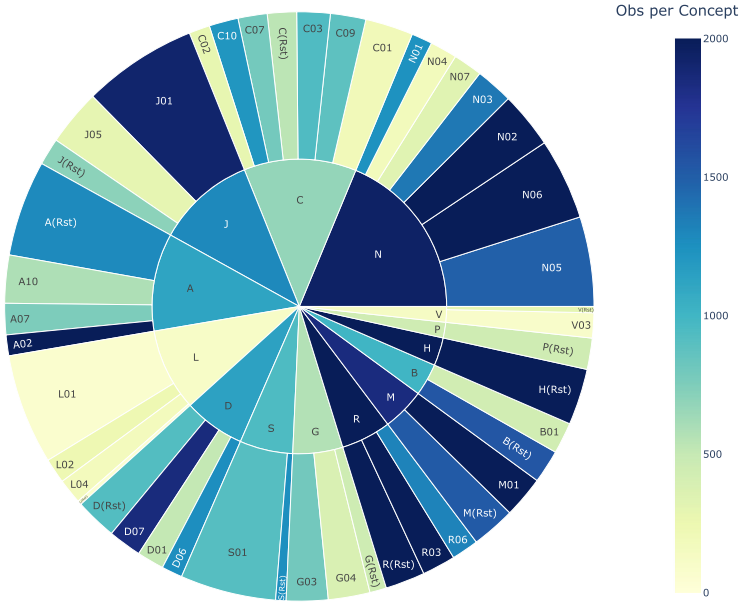
4.1 Data Description

We extract a large-scale depression patient cohort consisting of 37,669 patients during 2011 to 2014 from IBM MarketScan claims database[3] [2]. Specifically, following the depression cohort designed in Reference [19], patients are included if they had at least one exposure to an antidepressant medication with at lease one diagnosis code for depression disease. For each patient, we retrieve the diagnosis and the prescriptions spanning the records of inpatient, outpatient, and facility visits since 1 year before the depression index date (i.e., patient's first exposure to antidepression treatment) until the latest visit when the patient was prescribed a new antidepressant drug. We merge all these information from different sources of visits by patient ID and sort the information of each patient by the date of service incurred. For prescription observations, since we aim to predict the new prescriptions for the purpose of discovering treatment pathways for patients, we always consider the first-time prescription of a drug as the effective patient-drug observation for each patient. In total, we have 7,222 unique diagnosis and 875 unique drugs including 30 antidepressant drugs treating depression observed in data. For the ontology information. We have the ICD-9 Onto-DAG consisting of 8,248 diagnosis concepts, and the ATC Onto-DAG with 1,487 drug therapeutic concepts. More ontology details are shown in Figure 6, where the portion of each unique concept is visualized and their frequency of observation in the dataset is indicate by the color. Other than ontology, prescription, and diagnosis information, we collect demographic feature of patients and side-effect features of drugs. The demographic features including age and other information such as sex, employment status, geographic location, and so on. Through feature binarization, we obtain the demographic vector with 90 dimensions for each patient. For the side effect of drugs, we collect the **Medical Dictionary for Regulatory Activities Terminology (MedDRA)** showing the possible adverse reactions caused by drugs such as *Allergic conditions*, *Body temperature conditions*, and *Cardiac*

[3] https://www.ibm.com/products/marketscan-research-databases.



(a) ICD-9 ontology of diagnosis



(b) ATC ontology of drugs

Fig. 6. Distribution of unique concepts in ICD-9 and ATC ontologies of experimental dataset. Due to the large number of concepts at lower levels, we only visualize the two topmost levels. The inner layer shows the portion of concepts at the first level in terms of the number of unique diagnosis/drugs while the outer layer shows the distribution at the second level. The color indicates the frequency of ontology concepts in the dataset, which is normalized by the portion size (e.g., Observations per concept). Very small portions are combined and shown as "Rst."

Table 2. Data Statistics

MarketScan	
# of patients	37,669
# of drug codes	875
# of diagnosis codes	7,222
Age (mean ± sd)	45 ± 13
Sex (male/female)	10,275/27,394
# of patient-drug pairs	662,123
Avg drug per patient	17.58
Avg diagnosis per patient	25.87
# of nodes in ATC Onto-DAG	1,487
# of nodes in ICD Onto-DAG	8,248

arrhythmias. By counting all the unique side effects, we obtain the entire side-effect vector with 312 dimensions for each drug. Finally, by labeling the patient-drug pairs where 1 means observed prescription and 0 means negative sampled prescription, we have 662,123 positive patient-drug interactions and 10 times of them as negative interactions. More data statistics are shown in Table 2.

4.2 Baseline Methods

We compare the proposed OntoPath[4] framework with baselines including both recommendation methods and healthcare predictive methods.

- **SVD++** [22]: an extended **matrix factorization (MF)** model based on **singular value decomposition (SVD)**.
- **BPR** [33]: a top-k recommendation framework using the Bayesian personalized ranking optimization method for matrix factorization.
- **DeepFM** [14]: a deep learning model that models low- and high-order feature interactions introduced by factorization machines
- **NeuMF** [17]: a deep learning model with a **generalized matrix factorization (GMF)** and a MLP networks for modeling user-item interactions.
- **Med2Vec** [7]: a word2vec-based framework learning the embeddings of medical codes by predicting the code in the neighbor visits using the current visit. Once we obtain the trained diagnosis code embeddings, we train a recurrent neural network to predict the drug in the next visit.
- **RETAIN** [9]: an interpretable medical code predictive model based on two cooperative reversed time recurrent neural networks. For each patient-drug pair, we use the history diagnosis codes as input for predict the drug for the next visit.
- **GRAM** [8]: a medical ontology-based model for code prediction by supplementing diagnosis codes with hierarchical information using graph-based attentions. History diagnosis code along with ICD ontology is used for predicting next visit drugs.
- **G-Bert** [37]: a medication recommendation model that combines the **Graph Neural Network (GNN)** for ontology structure learning and **Bidirectional Encoder Representations from Transformers (BERT)** for sequential medical code pre-training.

[4]The code is available at: https://github.com/zyao237/ontopath.

- **HAP** [48]: a medical ontology embedding model that proposes to use **Hierarchical Attention Propagation (HAP)** for propagating attention across the entire ontology structure, learns medical concepts from not only ancestors but also descendants, siblings, and others.

4.3 Experimental Setting

In the training stage, we optimize the model with all the possible prescriptions (covering 875 unique drugs in total), to fully learn the medical characteristics of each patient by making model sophisticated and discriminative. In the evaluation stage, to mimic the practical scenario where finding next antidepressant treatment is the main target for depression treatment pathways, we narrow down the recommendation candidates to all the available antidepressant drug (30 unique drugs in total). To this end, we randomly sample 50% patients, to whom we adopt leave-one-out evaluation by withholding their last prescribed antidepressant drugs for testing and returning all the previous prescriptions for training. For each patient-drug prescription instance, we use the patient's history diagnosis before the prescription date, demographic and side-effect information, as well as the Onto-Paths of diagnosis codes and drugs as input. With the predicted probabilities, we rank all the candidate antidepressant drugs for each testing patient, and use top-k evaluation metrics to validate the accuracy of drugs in recommendation lists.

For implementation, we empirically set the following hyperparameters: all embedding size as 64, batch size as 256, negative sampling as 10, dropout as 0.2, initializer as normal, recurrent encoder of hierarchical concepts as GRU, attention heads of Transformer for diagnosis summarization as 2, similarity function of contrastive loss for pre-training as dot product, classification optimizer as SGD with 0.9 momentum, pre-training epoch as 5 then prescription recommendation training until converge.

4.4 Evaluation Metrics

For each testing patient, we have the last antidepressant prescription as the ground truth and the rest of the unprescribed drugs from the 30 total antidepressants as the negative candidates (testing patients have 27 negative antidepressants on average). By ranking the ground truth and negative antidepressants by predicted probabilities, we generate a top-k recommendation list for each patient and evaluate it with the two following metrics. Finally, the overall performance is reported by averaging the results over all the testing patients.

- **Hit Ratio (HR@k)**: given the top-k list for antidepressant recommendations, we check if the ground-truth drug is in the list. If yes, then we mark 1 for this patient and 0 otherwise.
- **Normalized Discounted Cumulative Gain (NDCG@k)**: given the top-k list for antidepressant recommendations, we consider the rank position of the ground-truth drug in the list. The score decrease as the ground truth's rank goes lower (0 if out of ranking list).

4.5 Performance Comparisons

We present the performance comparisons between the proposed OntoPath and the baseline approaches on prescription recommendation in Table 3. Both **Hit Ratio (HR)** and **Normalized Discounted Cumulative Grain (NDCG)** examine the ranking quality of drug recommendations. Due to the high-stakes decisions of treatment planning for depression patients, we use relatively short recommendation list for evaluation. For example, HR@1 means the average HR score over all testing patients that if the top-1 recommending drug identifies the ground-truth antidepressant. In total, we evaluate the recommendation list on five different sizes from 1 to 5.

Generally, we can see that the proposed approach OntoPath consistently outperforms other baselines on all metrics in each top-k setting. Specifically, first we can see that classic MF approach

Table 3. Performance Comparisons

Methods	Hit Ratio					NDCG				
Top-k	@1	@2	@3	@4	@5	@1	@2	@3	@4	@5
SVD++	0.2331	0.3649	0.4674	0.5666	0.6593	0.2331	0.3139	0.3648	0.4077	0.4440
BPR	0.2288	0.3465	0.4534	0.5605	0.6619	0.2288	0.3012	0.3541	0.4001	0.4399
DeepFM	0.2338	0.3555	0.4503	0.5574	0.6597	0.2338	0.3082	0.3555	0.4014	0.4415
NeuMF	0.2350	0.3606	0.4545	0.5489	0.6356	0.2350	0.3120	0.3586	0.3990	0.4325
Med2Vec	0.2290	0.3500	0.4561	0.5652	0.6660	0.2290	0.3027	0.3556	0.4027	0.4422
RETAIN	0.2362	0.3611	0.4727	0.5720	0.6609	0.2362	0.3115	0.3673	0.4104	0.4451
GRAM	0.2597	0.3887	0.4878	0.5825	0.6673	0.2597	0.3371	0.3867	0.4278	0.4610
G-Bert	0.2479	0.3830	0.4914	0.5857	0.6718	0.2479	0.3296	0.3839	0.4246	0.4584
HAP	0.2591	0.3847	0.4938	0.5899	0.6727	0.2591	0.3346	0.3893	0.4310	0.4633
OntoPath	**0.2671**	**0.3979**	**0.5092**	**0.6032**	**0.6834**	**0.2671**	**0.3460**	**0.4017**	**0.4425**	**0.4739**

SVD++ give very competitive performance comparing to other recommendation methods (BPR, DeepFM, NeuMF). Then, we see that the BPR model gives the worse performance. A potential reason is that since we use implicit feedback (1 or 0) in this study, BPR can be constrained by using the ranking order of each item pair as optimization objectives instead of classification losses. The deep learning recommendation approaches (DeepFM, NeuMF) do not show significant advantages comparing to traditional recommendation models. DeepFM modeling the higher order of feature interaction has better performance than BPR but loses to SVD++. NeuMF performs better than DeepFM but is still similar to SVD++. Finally, we have the medical predictive baselines (Med2Vec, RETAIN, GRAM, G-Bert, HAP), which start to use either diagnosis or ontology structure as supporting information. First, Med2Vec does not perform well as it is an unsupervised embedding framework. The diagnosis codes are trained based on the co-occurrence with neighbor visits instead of recommendation labels. Second, RETAIN achieves better overall performance than all the previous baselines, showing that diagnosis history is a necessary evidence to train a drug recommender, however it is not a personalized method and medical domain knowledge has not been introduced. Third, GRAM and G-Bert make the similar performances and provide better result than Med2Vec and RETAIN by starting to incorporate medical ontology information like propagating ancestor concepts information in medical concept representation. Last, as an improved model comparing to GRAM, HAP facilitates a more sophisticate hierarchical propagation for ontology structure learning, and achieves the best overall performance except OntoPath. By utilizing the entire neighborhood of target concepts through both bottom-up and top-down rounds of propagation with attention, HAP shows similar or better results than GRAM and G-Bert by generating comprehensive patient representation.

4.6 Ablation Study

To further validate the contribution of each component to the model, we conduct ablation study by removing a component at a time for the three contributions we claim on OntoPath. Table 4 shows the HR and NDCG performances for the three ablation cases. For *No Onto.*, we remove the model components of using ontology information of ICD-9 and ATC, but we still have the diagnosis sequence learning by Transformer with patient demographic and drug side-effect feature as auxiliary information. For *No Diag.*, we remove the component consuming diagnosis code and

Table 4. Ablation Study

Methods	Hit Ratio					NDCG				
Top-k	@1	@2	@3	@4	@5	@1	@2	@3	@4	@5
No Onto.	0.2435	0.3738	0.4834	0.5862	0.6708	0.2435	0.3224	0.3769	0.4216	0.4547
No Diag.	0.2299	0.3482	0.4602	0.5642	0.6608	0.2299	0.3027	0.3581	0.4030	0.4407
No Pre-tr.	0.2615	0.3881	0.4988	0.5943	0.6759	0.2615	0.3376	0.3932	0.4345	0.4665
Full	0.2671	0.3979	0.5092	0.6032	0.6834	0.2671	0.3460	0.4017	0.4425	0.4739

Table 5. Prescription Recommendation of ATC Concepts

Methods	Hit Ratio					NDCG				
Top-k	@1	@2	@3	@4	@5	@1	@2	@3	@4	@5
ATC lv-3	0.5842	0.9388	0.9442	0.9482	0.9524	0.5842	0.7057	0.7540	0.7810	0.7984
ATC lv-4	0.4033	0.6629	0.7938	0.8597	0.8928	0.4033	0.4611	0.4818	0.4901	0.4962
ATC lv-5	0.2671	0.3979	0.5092	0.6032	0.6834	0.2671	0.3460	0.4017	0.4425	0.4739

the auxiliary demographic information. Because of the removal of the entire diagnosis evidences, ICD-9 ontology is unused, but ATC ontology is still working. Last, we remove the pre-training stage as *No Pre-tr.* to show how effective the contrastive pre-training component is for enhancing the performance of final patient-drug prescription prediction.

From the results, we observe that comprehensive diagnosis evidences are the first contributor to drug recommender as *No Diag.* loses the largest margin of performance. Ontology information is the second contributor as *No Onto.* shows the second largest decline on metrics. Last, although pre-training does not enhance performance as large as the two previous components, it enhances the best performance of recommender considering OntoPath without pre-training already outperforms all the baselines.

4.7 Recommendation of ATC Concepts

We examine the performance of recommending more general ATC concepts instead of specific drugs. The idea is that two drugs may be different on the exact form of drug ingredients (at ATC level 5), but they may have the same drug properties (e.g., therapeutic functions) by sharing the same concepts at level 4 or level 3 of ATC ontology. It means that the predicted drugs, which share part of therapeutic functions with the ground-truth drug, can still be effective for treatment to some degree. Specifically, in this study, we use the higher level (e.g., level 3 or 4) ATC concepts of drugs as the new ground truth for drug recommendation. As a result, the drugs sharing the same ATC concept with the originally positive drug will be labeled positive as well. To validate the effectiveness of recommendation of ATC concepts, Table 5 shows the prediction results of recommending ATC concepts on level 3 and level 4 against the original evaluation on specific drugs (i.e., ATC level 5). We can see that the recommendation of drugs on ATC level 4 (lv-4) has increased the HR@1 from 0.2671 to 0.4033, which means that more recommended antidepressant drugs may be potentially effective as well. If we examine the recommendations at ATC level 3 (lv-3), then the HR@2 approaches to 0.9388. It makes the drug recommender relatively reliable on ATC level 3, but accordingly the drug effectiveness will be more uncertain, since the therapeutic functions sharing at ATC level 3 cover much more drug concepts.

Table 6. ICD-9 Codes with the Highest Attentions Learned for Predicting Example Antidepressant Drugs

Antidepressant	Bupropion	Trazodone	Sertraline
Diagnosis with the highest attention (ICD-9)	Hypothyroidism (244.9) Headache (784.0) Insomnia (780.52) Tobacco use disorder (305.1) Myalgia and myositis (729.1) Esophageal reflux (530.81) Cough (7862) Recur depr psych-severe (296.33) Anxiety state (300.00) Hyperlipidemia (272.4) Malaise and fatigue (780.79) Urin tract infection (599.0) Vitamin D deficiency (268.9) Nausea with vomiting (787.01)	Insomnia (780.52) Hypothyroidism (244.9) Headache (784.0) Myalgia and myositis (729.1) Tobacco use disorder (305.1) Recur depr psych-severe (296.33) Esophageal reflux (530.81) Cough (786.2) Anxiety state (300.00) Nausea with vomiting (787.01) Hyperlipidemia (272.4) Urin tract infection (599.0) Vitamin D deficiency (268.9) Diarrhea (787.91) Morbid obesity (278.01)	Hypothyroidism (244.9) Headache (784.0) Insomnia (780.52) Cough (786.2) Myalgia and myositis (729.1) Tobacco use disorder (305.1) Esophageal reflux (530.81) Anxiety state (300.00) Recur depr psych-severe (296.33) Hyperlipidemia (272.4) Urin tract infection (599.0) Nausea with vomiting (787.01) Pain in limb (729.5) Acute bronchitis (466.0)
Antidepressant	Olanzapine	Chlordiazepoxide	Fluvoxamine
Diagnosis with the highest attention (ICD-9)	Recur depr psych-severe (296.33) Headache (784.0) Insomnia (780.52) Hypothyroidism (244.9) Anxiety state (300.00) Tobacco use disorder (305.1) Psychosis (298.9) Cough (786.2) Esophageal reflux (530.81) Opioid dependence (304.00) Altered mental status (780.97) Rec depr psych-psychotic (296.34) Myalgia and myositis (729.1)	Irritable bowel syndrome (564.1) Hypothyroidism (244.9) Esophageal reflux (530.81) Tobacco use disorder (305.1) Insomnia (780.52) Myalgia and myositis (729.1) Abdmnal pain epigastric (789.06) Nausea with vomiting (787.01) Hyperlipidemia (272.4) Diarrhea (787.91) Recur depr psych-severe (296.33) Anxiety state (300.00) Abnormal loss of weight (783.21) Sympt fem climact state (627.2) Lump or mass in breast (611.72)	Recur depr psych-severe (296.33) Headache (784.0) Insomnia (780.52) Cough (786.2) Adjust dis w anxiety/dep (309.28) Anxiety state (300.00) Hypothyroidism (244.9) Hyperlipidemia (272.4) Lump or mass in breast (611.72) Tobacco use disorder (305.1) Fever (780.60) Rec depr psych-psychotic (296.34) Chronic pain (338.29) Esophageal reflux (530.81) Myalgia and myositis (729.1)

ICD-9 codes are ranked by the frequency of having the highest attentions in patients' medical history.

4.8 Attention across Diagnosis Codes

One of the benefits of OntoPath is that the Transformer-based diagnosis learning module puts attention to different diagnosis codes when predicting different drugs. We conduct an experiment on inspecting the attention for history diagnosis to show the explainability of OntoPath. Specifically, we retrieve all the prescription instances of a particular antidepressant, and extract the Transformer decoding attention scores for combining history ICD-9 diagnosis codes. For each prescription instance, we keep the top-3 ICD-9 codes with the highest attention scores, then we rank the ICD-9 codes by their frequencies of being counted in top-3 attentions across all the instances.

Table 6 shows the diagnosis codes that gain the highest attentions in the prediction of six frequently used antidepressant drugs. The order of codes in each cell is ranked by frequencies from high to low. Each diagnosis code ranking list is combined from the results of 2 Transformer attention heads. Generally, we find that most of the diagnosis are relevant to the symptoms or complications of depression or mental illness. First, Insomnia (780.52), Headache (784.0), and Hypothyroidism (244.9) have been well-known symptoms of depression [21]. Moreover, Tobacco use disorder (305.1) and Opioid dependence (304.00) have been strongly related to antidepressant (e.g., Chlordiazepoxide) treatment reflecting the common experience of smoking cessation and Opioid withdrawal for depression patients [31]. Third, Sympt fem climact state (627.2) is highlighted as antidepressants are frequently used for treating patients with menopausal symptoms accompanied by anxiety [29]. Last, Vitamin D deficiency (268.9) associated with seasonal affective disorder can be an important factor causing depression [13].

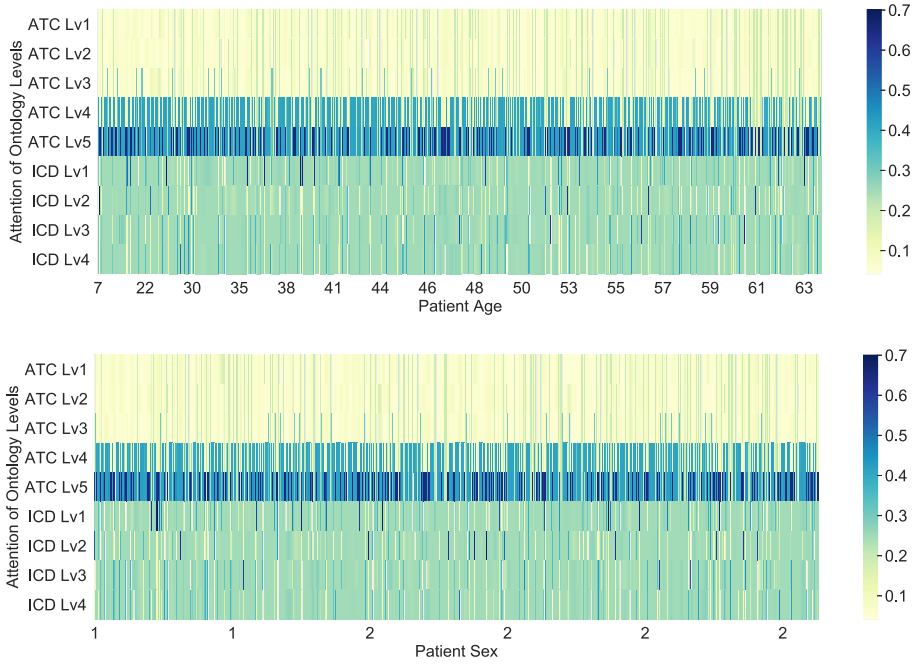


Fig. 7. Hierarchical attention distribution across ATC and ICD-9 ontology levels of randomly sampled patient-drug instances ordered by patient age or by sex (1: male, 2: female).

4.9 Attention across Ontology Levels

The hierarchical interaction between the ontology concepts of patients and drugs is coordinated by a co-attention max pooling network in OntoPath. Facing different patient-drug pairs, the attention network automatically adjust the attention on different ontology levels for extracting the most helpful information. To inspect how the model distributes attentions across different ontology levels, we visualize the attentions of individual cases. Figure 7 shows the attention heatmaps across both ATC levels and ICD-9 levels from 1,000 randomly sampled patient-drug pairs, where each column shows a specific patient-drug interaction. In addition, we add demographic information and rank the patients by two features: age and sex (1: male, 2: female). We can see that generally ATC attentions are more concentrated on level 4 and level 5, occasionally on level 3. However, ICD-9 attentions are more evenly distributed on all four levels. A possible reason is that each level of patient hierarchical embedding is uniquely extracted from a sequence of ICD-9 ontology concepts, which means that there are always unique information provided at each ontology level of a patient's diagnosis information.

4.10 Treatment Recommendations across Patient Subgroups

To further understand how the trained recommendation framework predicts treatments given the heterogeneity of patient characteristics, we categorize the testing patients into multiple subgroups and study the distribution of treatment recommendations, which lead to different treatment pathways. Figures 8 and 9 show the patient subgroups based on either demographic or phenotypic information, where the distribution of treatment recommendations are shown in heatmaps. To better illustrate the transition patterns of prescription happened in treatment pathways, we format the prescription recommendations as "last treatment → new treatment."

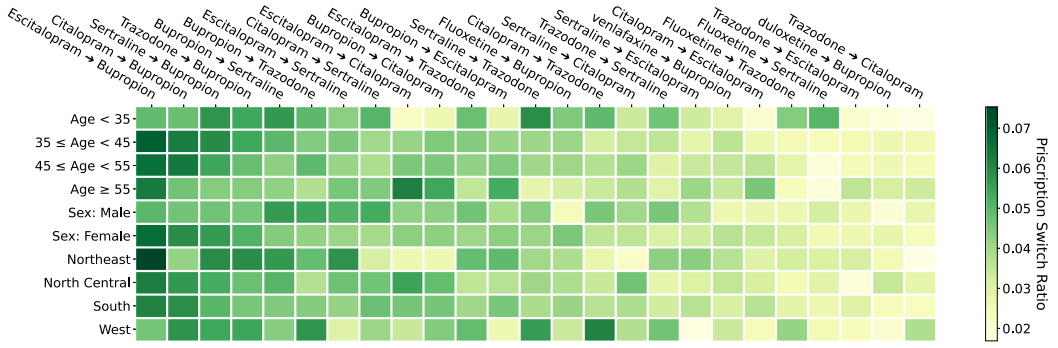


Fig. 8. The treatment pathways recommended for different patient subgroup by demographic information.

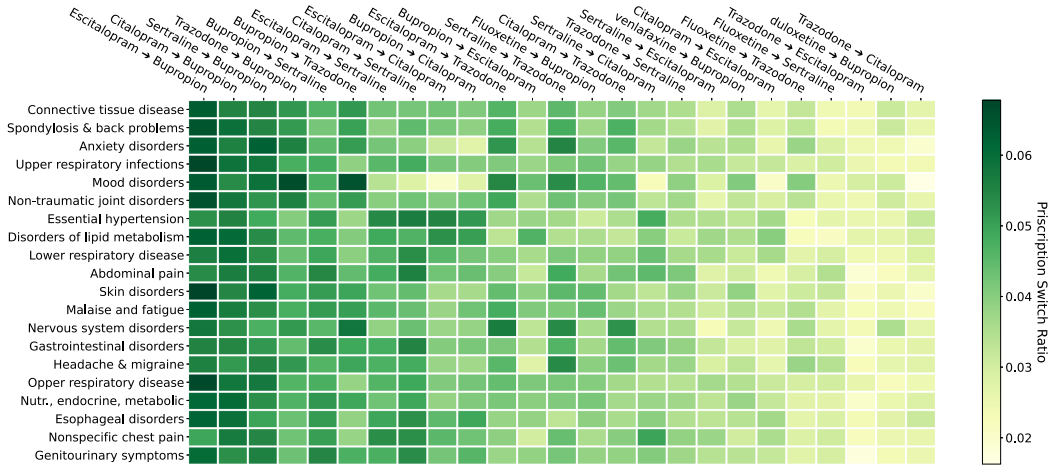


Fig. 9. The treatment pathways recommended for different patient subgroup by clinical phenotypes.

For example, color block at first row and first column in Figure 8 indicates the ratio of patients who were on *Escitalopram* antidepressant treatment, are recommended to have *Bupropion* as the new prescription, among all the depression patients in the subgroup of "younger than 35." We select the 25 most frequent prescription switch patterns resulted in top-2 recommendations in OntoPath, rank them from left to right in descending order, and normalize the frequencies of all the prescription recommendation as the ratio numbers between 0 and 1.

In Figure 8, we study the demographic patient subgroups in three categories—age, sex, and region. First, for the four age subgroups, we can see that the mid-age groups "$35 \leq$ Age < 45" and "$45 \leq$ Age < 55" have a similar patterns of recommendation, but the young group "Age < 35" and the elder group "Age ≥ 55" provide more distinction. For example, the transition pattern "*Sertraline* $\rightarrow$ *Trazodone*" and "*Fluoxetine* $\rightarrow$ *Sertraline*" are not very popular for the other age groups but are top choices for young patients. This differentiation of prescription recommendation reflects the model's capacity in treatment personalization as the cause and progression of depression in young and elder subgroups are significantly different than the other cohorts. Second, the two sex groups shows the similar distribution in general. Minor difference exists in the popular treatment pathways, such as "*Escitalopram* $\rightarrow$ *Bupropion.*" Last, by categorizing patients in the U.S. by four major regions, we observe some interesting patterns in the "west" and the "northeast" subgroups.

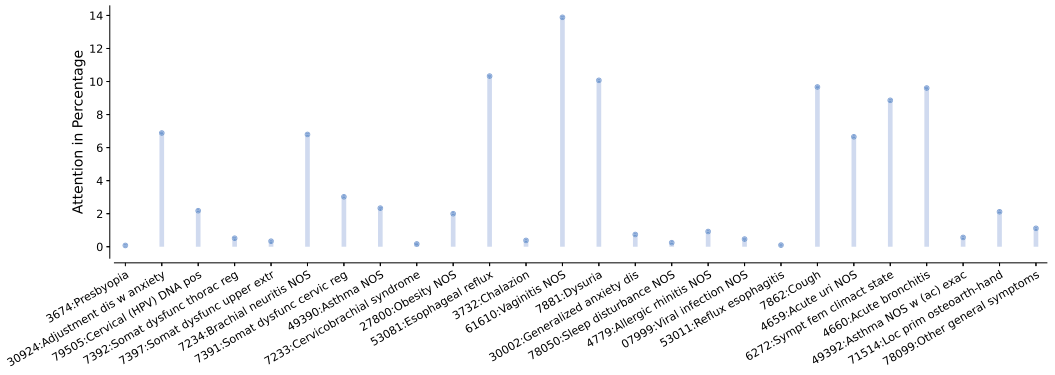


Fig. 10. Attention distribution over historical diagnosis of patient case 1.

For example, we can see that the patients from the "west" region have a notable difference in prescription recommendation comparing to the patients in other regions. For example, "*Citalopram* → *Trazodone*" is the most popular prescription switch pattern for west region patients but is not for the others. This differentiation in recommendations reveals that the prescription differences exist widely in observational EHR data by different areas, and suggests that the geolocation can be an influential factor in building clinical predictive applications.

In Figure 9, we study how the prescription recommendation changes if a patient is from different clinical phenotypic groups. To achieve this goal, we utilize the **Healthcare Cost and Utilization Project (HCUP)** database from the **Clinical Classification Software (CCS)** [12] to group ICD-9 codes into multiple homogeneous disease categories. We have ICD-9 codes of each testing patient go through the single-level categorization in CCS, and label each patient with any involved categories. Among all the possible disease categories, we pick the 20 most frequent disease categories as the highly relevant clinical phenotypes of depression patients. In details, we can see that patients with *Mood disorders*, *Essential hypertension*, and *Nervous system disorders* tend to have different prescription recommendations in their treatment pathways. For example, the prescription transition patterns "*Escitalopram* → *Trazodone*," "*Sertraline* → *Trazodone*," and "*Citalopram* → *Trazodone*" are more heavily used in patients with *Nervous system disorders* comparing with other phenotypes.

4.11 Attention Behavior on Individual Patients

In this section, we visualize the attention across the diagnosis codes of individual patients, to study how OntoPath learns the patient medical history through the tranformer-based attention modeling. In Figures 10 and 11, the attention scores on historical diagnosis codes of two individual patients are shown, who are randomly selected with as least 25 diagnosis codes and have hit the ground-truth prescription with top-2 recommendations. In each figure, the x-axis shows the ICD-9 diagnosis codes sorted in ascending order by time, and the y-axis shows the single head attention scores over each diagnosis in percentage.

Figure 10 shows the case study of a 49-year-old female patient who was on the antidepressant drug *Citalopram*, is recommended switching the prescription to *Bupropion*. We can see that this patient was having stress-related conditions (i.e., *Adjustment dis w anxiety*) in the past, which is highlighted with the attention around 7%. Also, her underlying health conditions also show impact on the prescribing decisions including the nervous system (i.e., *Brachial neuritis*), digestive system (i.e., *Esophageal reflux*), respiratory system (i.e., *Cough*, *Acute uri*, *Acute bronchitis*), and female genitourinary system (i.e., *Vaginitis*, *Dysuria*, *Sympt fem climact state*), which are highlighted with

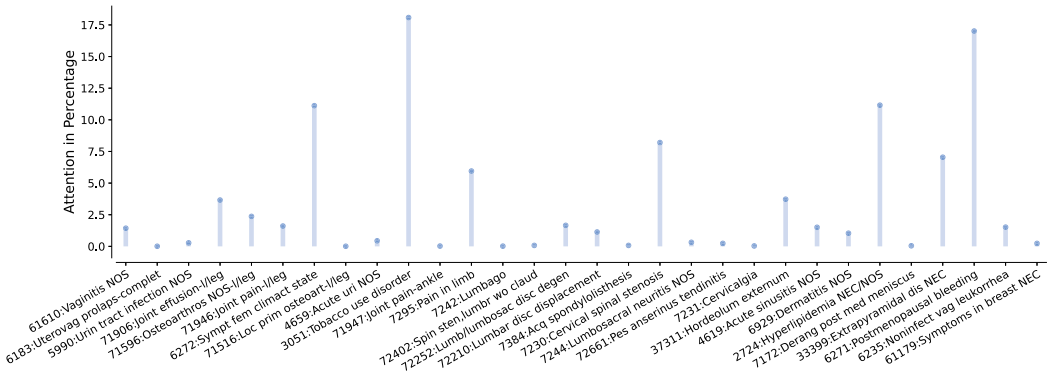


Fig. 11. Attention distribution over historical diagnosis of patient case 2.

relatively strong attentions as well. On the contrary, the irrelevant conditions such as the vision system (i.e., *Presbyopia*, *Chalazion*) have not attracted a good amount of attention from OntoPath.

Figure 11 shows the case study of a 56-year-old female patient who was on the antidepressant drug *Amitriptyline*, is recommended switching the prescription to *Duloxetine*. We can see that this patient was having a history of substance abuse condition (i.e., *Tobacco use disorder*), which is usually correlated with depression onset and is highlighted with the attention around 17%. Besides the major factor, her underlying health conditions concentrate on the pain and movement disorders (i.e., *Pain in limb*, *Cervical spinal stenosis*, *Extrapyramidal dis*), female menopausal disorders (i.e., *Postmenopausal bleeding*, *Sympt fem climact state*), and lipid metabolism disorders (i.e., *Hyperlipidemia*), which strongly influence the antidepressant recommendation through the attention scores.

5 RELATED WORK

Our work is related to two research areas: personalized recommender systems and healthcare predictive analytics.

5.1 Recommender Systems

Recommender systems have been successfully applied in various web services to help users find their interested information, such as E-commerce, social networks, and digital media. **Collaborative filtering (CF)** techniques profile user interests from observed user-item interactions without the requirement of domain knowledge, which have shown advantages of both accuracy and serendipity in real-life applications (e.g., Netflix). Basic CF algorithms include matrix factorization [23], Bayesian personalized ranking [33], factorization machines [32], and their variants [22, 28]. Recent years have witnessed an increasing number of work on applying deep learning techniques to recommender systems [49]. For example, NeuMF [17] combines the generalized matrix factorization and feed-forward MLP network to model implicit feedback data. DeepFM [14] derives an end-to-end model for capturing the low and high order feature interactions in a neural network architecture. The latest progress in Recommender systems involves employing graph data including ontology for enhancing recommendation accuracy. For example, NGCF [42] formulates the user-item interactions in a bipartite graph structure, and learns the embeddings by propagating information with high-order connectivity. KGNN-LS [41] transforms the knowledge graph into a user-specific graph and applies a graph convolutional network to compute the personalized embeddings.

Meanwhile, with the increasing demands of session-based applications (e.g., Youtube, Tiktok), sequential modeling with temporal consideration has been more and more introduced for next-item recommendation. Traditional sequential modeling are primarily in probabilistic framework such as **Markov Decision Processes (MDPs)** using transition probabilities [16, 34]. For example, FPMC [34] presents a matrix factorization method to learn a personalized transition graphs for each user to model sequential behaviors. In recently years, as deep learning techniques were proven to be more effective and flexible for capturing the dynamic of sequential information, RNN and its variants are increasingly used in literature [18, 20, 44, 51]. For example, Reference [18] adapts a GRU model to session-based recommendation by introducing a pair-wise ranking loss function. Reference [51] improves the LSTM network by equipping LSTM with time gates to accommodate the time interval information into sequential user behavior modeling. Reference [20] develops a GRU-based recommendation model with the incorporation of the **Key-value Memory Network (KV-MN)**, which is capable of leveraging knowledge base information to profile attribute-level user preferences.

5.2 Healthcare Predictive Analytics with Longitudinal Patient Records

In the domain of medical informatics, our work falls into the techniques that applies predictive analytic models and use longitudinal patient records to improve the healthcare outcomes. With the increasing availability of EHR and medical claims data, building predictive models from those data has attracted significant attention from both academia and industry.

Predicting disease diagnosis has been an active research focus. For example, RETAIN [9], Dipole [26], DoctorAI [6], and BRITS [4] present the RNN-based models with attention mechanisms aiming to predict the ICD codes by learning the influential visits and the key diagnosis in the past. GCT [11] proposes a Transformer graph neural network to learn the underlying structure of in-visit medical concepts by self-attention mechanism regularized by prior knowledge. HiTANet [25] proposes a time-aware self-attention network to embed time information to recognize the key timestamps in patient history. Besides the diagnosis code, the medication code is also a frequent target for prediction using longitudinal patient records. For example, LEAP [50] formulates the medication recommendation as a sequential prediction problem using a multi-instance (diagnosis) multi-label (drugs) learning framework by mapping a set of diagnosis with a set of drugs. G-Bert [37] proposes a step-by-step framework, which first uses graph neural networks to embed the medical concepts within ontologies, then adopts BERT model to pre-train medical code sequences, and finally builds a predictive model for medication recommendation. Gamenet [38] aims to recommend medication combination by integrating the **drug-drug interactions (DDI)** graph into the longitudinal patient record analysis. MedPath [46] learns a **personalized knowledge graph (PKG)** containing the possible disease progression paths from observed symptoms to target diseases, which are used to augment the EHR encoders for achieving better predictions. In addition to the diagnosis and medication prediction, other adverse health events have been also explored, such as hospital readmission prediction [15], mortality prediction [30], and so on.

Another topic that receives heavy research interests has been healthcare representation learning aiming to discriminatively describe patients for better facilitating disease progression modeling [1], and more downstream predictive tasks. For example, Med2Vec [7] learns the distributed representations of diagnosis codes and visits using the co-occurrence of codes in same visits and the sequential orders of visits in EHR data. Reference [43] proposes a continuous-time Markov process-based model to learn disease progression of chronic diseases in an unsupervised manner. Reference [24] uses temporal graphs to learn representation by capturing temporal relationships of medical events in EHR sequences. Metacare++ [39] proposes a meta-learning framework to address cold-start issues by developing a hierarchical patient subtyping strategy to bridge the

modeling of infrequent patients and rare diseases. Reference [45] develops a GNN-based model with time-aware meta-paths and self-attention mechanism to extract temporal semantics and inherent relations to learn an effective representation.

The last focus that has been extensively explored recently is to leverage domain knowledge like hierarchical ontology concepts for driving reliable and interpretable models or representations. For example, GRAM [8], HAP [48], and MiME [10] develop graph-base attention models that represent medical concepts with integration of structure information (e.g., ancestors) from medical ontologies. KAME [27] proposes a future visit prediction model that explicitly makes use of medical knowledge in the whole prediction process. Reference [47] proposes a **domain knowledge guided recurrent neural networks (DG-RNN)** by introducing medical knowledge graph into RNN architecture, as well as taking the irregular time intervals into account. G-Bert [37] and HyperCore [3] belong to this category as well.

There are several factors distinguish our study from the literature. First, unlike the previous medication recommendation work, which mostly adopts a drug code prediction task [37, 50], our study adopts a classic recommender systems problem setting aiming to recommend new drugs that have not been used by the patient before, instead of predicting the drugs that can be frequently used and repeatable (e.g., refill) in doctor visits. Second, unlike most of the longitudinal prediction work, which develop an universal model in a one-size-fits-all manner to predict for all patients [9, 26], our study targets at developing personalized model by profiling patients with unique representations, which is a critical step to achieve quality recommendations of prescriptions. Last, comparing to previous work that focus on the medical concept representation [7, 48], or adopt a stepwise framework for ontology and medical history learning [37], our model is an end-to-end trained framework integrating the learning of ontology information, longitudinal patient records, and final drug recommendation in a unified process.

6 CONCLUSION

In this article, we developed a multi-evidence prescription recommendation framework to leverage medical ontology information, personal diagnosis history, and auxiliary demographic and side-effect information, to discover effective drugs for chronic disease patients in treatment pathways. Specifically, we first incorporated a customized Transformer network to learn the sequences of ICD-9 diagnosis concepts extracted from patients' personal diagnosis history and further summarized the overall patient conditions using demographic information to achieve a comprehensive patient profiling. Second, we developed a dual-RNN encoder for ontology structure processing and a co-attention network to hierarchically model the level-to-level interactions for each patient-drug interaction. By using domain-specific ICD-9 concepts and ATC concepts for representing patients and drugs, respectively, we managed to model an in-depth patient-drug relevance to guide the decision making for prescription. Last, we exploited a contrastive loss to pre-train the patient and drug discriminativeness, to prepare a premium model initialization for boosting the final prescription predictive learning. We conducted extensive experiments on a real-world patient claims database of a large-scale depression patient cohort. The results showed that OntoPath outperformed all the baselines in terms of prescription recommendation performance, and the effectiveness of OntoPath was further validated with model interpretation in case studies.

REFERENCES

[1] Tian Bai, Shanshan Zhang, Brian L. Egleston, and Slobodan Vucetic. 2018. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 43–51.

[2] Anne M. Butler, Katelin B. Nickel, Robert A. Overman, and M. Alan Brookhart. 2021. IBM MarketScan research databases. In *Databases for Pharmacoepidemiological Research*. Springer, 243–251.
[3] Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3105–3114.
[4] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. 2018. Brits: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*. MIT Press, 6775–6785.
[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1597–1607.
[6] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. Doctor AI: Predicting clinical events via recurrent neural networks. In *Proceedings of the Machine Learning for Healthcare Conference*. 301–318.
[7] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1495–1504.
[8] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 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.
[9] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*. MIT Press, 3504–3512.
[10] Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. 2018. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *Advances in Neural Information Processing Systems*. MIT Press, 4547–4557.
[11] Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. 2020. Learning the graphical structure of electronic health records with graph convolutional transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 606–613.
[12] Anne Elixhauser, Claudia Steiner, and L. Palmer. 2015. Clinical classifications software (CCS). U.S. Agency for Healthcare Research and Quality. Available: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/CCSUsersGuide.pdf.
[13] F. M. Gloth 3rd, Waheed Alam, and Bruce Hollis. 1999. Vitamin D vs broad spectrum phototherapy in the treatment of seasonal affective disorder. *J. Nutr. Health Aging* 3, 1 (1999), 5–7.
[14] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A factorization-machine based neural network for CTR prediction. Retrieved from https://arXiv:1703.04247.
[15] Danning He, Simon C. Mathews, Anthony N. Kalloo, and Susan Hutfless. 2014. Mining high-dimensional administrative claims data to predict early hospital readmissions. *J. Amer. Med. Info. Assoc.* 21, 2 (2014), 272–279.
[16] Ruining He and Julian McAuley. 2016. Fusing similarity models with markov chains for sparse sequential recommendation. In *Proceedings of the IEEE 16th International Conference on Data Mining (ICDM'16)*. IEEE, 191–200.
[17] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*. 173–182.
[18] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations*.
[19] George Hripcsak, Patrick B. Ryan, Jon D. Duke, Nigam H. Shah, Rae Woong Park, Vojtech Huser, Marc A. Suchard, Martijn J. Schuemie, Frank J. DeFalco, Adler Perotte, et al. 2016. Characterizing treatment pathways at scale using the OHDSI network. *Proc. Natl. Acad. Sci. U.S.A.* 113, 27 (2016), 7329–7336.
[20] Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y. Chang. 2018. Improving sequential recommendation with knowledge-enhanced memory networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. 505–514.
[21] Sidney H. Kennedy. 2008. Core symptoms of major depressive disorder: Relevance to diagnosis and treatment. *Dialog. Clin. Neurosci.* 10, 3 (2008), 271.
[22] Yehuda Koren. 2008. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 426–434.
[23] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
[24] Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. 2015. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 705–714.

[25] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 2020. 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 and Data Mining*. 647–656.

[26] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 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.

[27] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. 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.

[28] Andriy Mnih and Russ R. Salakhutdinov. 2008. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*. MIT Press, 1257–1264.

[29] Stephanie Mulhall, Ross Andel, and Kaarin J. Anstey. 2018. Variation in symptoms of depression and anxiety in midlife women by menopausal status. *Maturitas* 108 (2018), 7–12.

[30] Nozomi Nori, Hisashi Kashima, Kazuto Yamashita, Hiroshi Ikai, and Yuichi Imanaka. 2015. Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 855–864.

[31] IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, International Agency for Research on Cancer, and World Health Organization. 1999. *Hormonal Contraception and Post-menopausal Hormonal Therapy*. Vol. 72. World Health Organization.

[32] D. L. Sackett, W. M. Rosenberg, J. M. Gray, R. B. Haynes, and W. S. Richardson. 1996. Evidence based medicine: what it is and what it isn't. *BMJ* 312, 7023 (1996), 71–72. Available: https://www.bmj.com/content/312/7023/71.

[33] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. Retrieved from https://arXiv:1205.2618.

[34] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized Markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web*. 811–820.

[35] David L. Sackett, William M. C. Rosenberg, J. A. Muir Gray, R. Brian Haynes, and W. Scott Richardson. 1996. Evidence-based medicine: What it is and what it isn't.

[36] Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive pooling networks. Retrieved from https://arXiv:1602.03609.

[37] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*. 5953–5959.

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

[39] Yanchao Tan, Carl Yang, Xiangyu Wei, Chaochao Chen, Weiming Liu, Longfei Li, Jun Zhou, and Xiaolin Zheng. 2022. MetaCare++: Meta-learning with hierarchical subtyping for cold-start diagnosis prediction in healthcare data. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'22)*. 449–459.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. MIT Press, 5998–6008.

[41] Hongwei Wang, Fuzheng Zhang, Mengdi Zhang, Jure Leskovec, Miao Zhao, Wenjie Li, and Zhongyuan Wang. 2019. Knowledge-aware graph neural networks with label smoothness regularization for recommender systems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 968–977.

[42] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 165–174.

[43] Xiang Wang, David Sontag, and Fei Wang. 2014. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 85–94.

[44] Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alexander J. Smola, and How Jing. 2017. Recurrent recommender networks. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*. 495–503.

[45] Yuyang Xu, Haochao Ying, Siyi Qian, Fuzhen Zhuang, Xiao Zhang, Deqing Wang, Jian Wu, and Hui Xiong. 2022. Time-aware context-gated graph attention network for clinical risk prediction. *IEEE Trans. Knowl. Data Eng.* (2022).

[46] Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. Medpath: Augmenting health risk prediction via medical knowledge paths. In *Proceedings of the Web Conference 2021*. 1397–1409.

[47] Changchang Yin, Rongjian Zhao, Buyue Qian, Xin Lv, and Ping Zhang. 2019. Domain knowledge guided deep learning with electronic health records. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'19)*. IEEE, 738–747.

[48] Muhan Zhang, Christopher R. King, Michael Avidan, and Yixin Chen. 2020. Hierarchical attention propagation for healthcare representation learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 249–256.

[49] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surveys* 52, 1 (2019), 1–38.

[50] Yutao Zhang, Robert Chen, Jie Tang, Walter F. Stewart, and Jimeng Sun. 2017. LEAP: Learning to prescribe effective and safe treatment combinations for multimorbidity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1315–1324.

[51] Yu Zhu, Hao Li, Yikang Liao, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. 2017. What to do next: Modeling user behaviors by time-LSTM. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*, Vol. 17. 3602–3608.

Received 21 June 2022; revised 5 November 2022; accepted 11 December 2022