MetaCare++: Meta-Learning with Hierarchical Subtyping for Cold-Start Diagnosis Prediction in Healthcare Data

Yanchao Tan  
College of Computer Science, Zhejiang University, China  
yctan@zju.edu.cn

Carl Yang  
Department of Computer Science, Emory University, USA  
j.carlyang@emory.edu

Xiangyu Wei  
College of Computer Science, Zhejiang University, China  
weixy@zju.edu.cn

Chaochao Chen  
College of Computer Science, Zhejiang University, China  
zjuccc@zju.edu.cn

Weiming Liu  
College of Computer Science, Zhejiang University, China  
21831010@zju.edu.cn

Longfei Li  
Ant Group  
China  
longyao.llf@antgroup.com

Jun Zhou  
Ant Group  
China  
jun.zhoujun@antfin.com

Xiaolin Zheng  
College of Computer Science, Zhejiang University, China  
xzheng@zju.edu.cn

ABSTRACT

Cold-start diagnosis prediction is a challenging task for AI in healthcare, where often only a few visits per patient and a few observations per disease can be exploited. Although meta-learning is widely adopted to address the data sparsity problem in general domains, directly applying it to healthcare data is less effective, since it is unclear how to capture both the temporal relations in clinical visits and the complicated relations among syndromic diseases for precise personalized diagnosis. To this end, we first propose a novel Meta-learning framework for cold-start diagnosis prediction in healthCare data (MetaCare). By explicitly encoding the effects of disease progress over time as a generalization prior, MetaCare dynamically predicts future diagnosis and timestamp for infrequent patients. Then, to model complicated relations among rare diseases, we propose to utilize domain knowledge of hierarchical relations among diseases, and further perform diagnosis subtyping to mine the latent syndromic relations among diseases. Finally, to tailor the generic meta-learning framework with personalized parameters, we design a hierarchical patient subtyping mechanism and bridge the modeling of both infrequent patients and rare diseases. We term the joint model as MetaCare++. Extensive experiments on two real-world benchmark datasets show significant performance gains brought by MetaCare++, yielding average improvements of 7.71% for diagnosis prediction and 13.94% for diagnosis time prediction over the state-of-the-art baselines.

CCS CONCEPTS

- Applied computing → Life and medical sciences.

KEYWORDS

Diagnosis prediction, Meta-learning, Hierarchical, Subtyping

ACM Reference Format:

1 INTRODUCTION

Electronic health records (EHRs) consisting of patients’ temporal visit information enable researchers and doctors to build better predictive models for clinical decision making [6, 20]. Many recent studies on this problem leverage modern deep learning models, such as recurrent neural networks [6, 22], attention-based mechanisms [23], and graph neural networks [7]. These models typically work well when adequate EHR data with task-specific labels are available, but can seriously suffer when training data are scarce [41, 43], for example, facing infrequent patients (i.e., patients with only a few visits) and rare diseases (i.e., diseases with only a few observations). In particular, we term the diagnosis prediction that involves infrequent patients and/or rare diseases as cold-start diagnosis prediction.

Recently, meta-learning has been demonstrated as an effective mechanism to alleviate the data sparsity problem in general domains (e.g., computer vision [35, 39], natural language processing [26, 42], and recommendation [8, 16]). Although these models encode general knowledge from huge amounts of training data, they ignore the specific modeling of healthcare data, such as the temporal relations between patients’ sequential diagnoses [20] and complicated relations among syndromic diseases [23]. For example, as shown in Figure 1, Jack is an infrequent patient with only two historical visits, who suffers from a high blood pressure disease (i.e.,
Malignant Essential Hypertension (401.0) at the diagnosis time $T_3$. To predict such a rare disease that only appears a few times in the whole database, an ideal meta-learner should (1) consider the accumulative effect of disease progression over the individual patient, such as Abnormal Glucose (790.2) from $T_1$, since a significant period of time with 790.2 will increase the risk of high blood pressure; (2) consider the hierarchical relations among diseases that categorizes 401.0 as a sub-disease of high blood pressure and the syndromic relations between high blood pressure and abnormal glucose.

In this work, we enable such a specialized clinical meta-learning framework to achieve precise diagnosis prediction. The task is challenging from several perspectives.

**Challenge I: How to design a meta-learner to capture temporal relations from patients’ sequential visits?** In healthcare data, since diseases can progress over time [29, 30, 44], the diagnosis time and the temporal relations among each patient’s sequential diagnoses (as shown in Figure 1) are important for diagnosis prediction. Although existing meta-learning methods that can be divided into the support set and query set can derive the global knowledge from sets as a generalization prior [9, 16, 33], such permutation-invariant sets fail to model the above accumulative effects of disease progression among sequential diagnoses.

**Challenge II: How to model both the hierarchical and syndromic relations among diseases?** To accurately model diseases, especially rare diseases that only appear a few times, it is practical to borrow the disease hierarchies (e.g., ICD-9 [2, 32]) from domain knowledge. By classifying diseases into various types according to body systems with a tree-structured hierarchy, we can allow the parent diseases to summarize common properties of child diseases, while the child diseases can inherit important properties from the parents. For example, as shown in Figure 1, 272.0 and 272.1 are children of 272, whereas 401.0 and 401.1 are children of 401. However, the existing ICD-9 hierarchy does not incorporate the syndromic relations among diseases (e.g., 272.0 in the metabolic system is easily concurrent with 401.0 in the circulatory system), which can be extracted from actual diagnosis data and valuable for diagnosis prediction.

**Challenge III: How to properly parameterize meta-learner for personalized diagnosis prediction?** The leading causes of one disease for different patients vary a lot [41]. However, a generic meta-learner with globally shared knowledge can backfire on personalized diagnosis prediction. How to fully leverage temporal relations among sequential diagnoses and complicated relations among syndromic diseases to personalize the meta-learner remains unknown.

To address the above challenges, we first propose MetaCare, which introduces a clinical meta-learner to capture temporal relations among patient visits for cold-start diagnosis prediction. Furthermore, we propose a diagnosis-enhanced disease representation learning method to model both the hierarchical and syndromic relations among diseases. Finally, we propose MetaCare++, which devises a hierarchical patient subtyping strategy to bridge the modeling of infrequent patients and rare diseases, and tailor the parameters of the meta-learner for personalized diagnosis prediction.

Our overall contributions in this work are summarized as follows:

- **Formulation of cold-start diagnosis prediction.** MetaCare++ is the first diagnosis prediction model for both infrequent patients and rare diseases, which can effectively alleviate the data sparsity problem in diagnosis prediction. (Section 3.1).
- **Effective model designs.** In MetaCare, we propose a novel clinical meta-learner, which captures the temporal relations among patients’ sequential diagnoses regarding both diagnoses and times (Section 3). Furthermore, in MetaCare++, we devise a diagnosis-enhanced disease representation learning and a personalized decoder via hierarchical subtyping, which captures both temporal relations among sequential diagnoses and complicated relations among syndromic diseases to achieve precise diagnosis prediction (Section 4).
- **Extensive experiments on real benchmark EHR datasets.** We conduct comprehensive experimental evaluations on cold-start diagnosis prediction tasks against state-of-the-art approaches over two public large-scale EHR datasets. Extensive experimental results demonstrate the superiority of MetaCare++ (Section 5).

2 RELATED WORK

2.1 Deep Learning for Diagnosis Prediction

Diagnosis prediction is a developing area that leverages a patient’s temporal visits to predict future diagnosis [6, 14, 19, 37, 41]. For example, RETAIN [6] employed an attention process on recurrent neural networks (RNNs) to model the order of visits for the disease prediction task. Dipole [22] applied bidirectional long-short term memory networks and attention mechanisms to predict patient visit information. Both Timeline [2] and ConCare [24] utilized
time-aware attention mechanisms in RNN for health event predictions. However, in the cold-start diagnosis prediction setting which involves infrequent patients and/or rare diseases, the above approaches suffer from the data sparsity problem [7, 27].

To alleviate the data scarcity challenge in healthcare, many works [7, 20, 23, 41] exploited the information from external medical knowledge graph for robust representations. For example, GRAM [7] constructed a disease graph from medical knowledge (KG). MedPath [41] used a personalized knowledge graph extracted from SemMed to learn the disease progression information for individual patient. GCL [20] utilized the hierarchical structure of medical domain knowledge and introduced an ontology weight to capture hidden disease correlations. However, existing medical KGs do not fully capture the complicated relations among diseases, and it is unclear how they can be used to model infrequent patients.

2.2 Meta Learning
Meta-learning has attracted tremendous attention due to its effectiveness in many domains, such as computer vision [35, 39], natural language processing [26, 42], and recommendation [8, 16]. Among them, optimization-based meta-learning is widely adopted, where a gradient procedure is trained to be applied on a learner directly [1, 10, 18]. For example, model-agnostic meta-learning (MAML) [10] aimed to learn a good parameter initialization for the fast adaptation of testing tasks. Based on MAML, MAMO [8] designed task-specific and feature-specific memory matrices for user cold-start and item cold-start problems in recommendation. TuNP [18] associated each user with a corresponding stochastic process and learned a task-specific meta-learning framework to enhance user cold-start recommendation.

While the advantages of meta-learning seem eminent, the application of meta-learning in healthcare has rarely been explored. Inspired by the above promising works based on meta-learning, we propose to follow a manner of parameter initialization for cold-start diagnosis prediction. As closest to us, [43] proposed to leverage labeled patients from other relevant high-resource domains under a multi-domain setting. [36] designed a task-adaptive meta-learning framework for solving the rare diseases problem. However, both of them fail to incorporate the temporal relations among sequential diagnoses to learn generalizations priors that facilitate the modeling of infrequent patients.

3 THE METACARE FRAMEWORK

3.1 Problem Statement
Our goal is to provide precise diagnosis prediction involving rare diseases and infrequent patients. We formulate the cold-start diagnosis prediction from the meta-learning perspective.

We first denote the diagnosis dataset as \( D \). For each patient \( u_i \), \( D \) includes a sequence of diagnosis \( \tau_i = \{d_{i,j}, T_{i,j}\}_{j=1}^{N_i} \), where \( d_{i,j} \) denotes the \( j \)-th diagnosed disease, \( T_{i,j} \) is the corresponding diagnosis time, and \( N_i \) denotes the number of diseases for \( u_i \). Then, we can divide each sequence of diagnosis into a support sequence \( S_i \) and a query sequence \( Q_i \) according the diagnosis time (\( \tau_i = S_i \cup Q_i \)). Thus, the task of cold-start diagnosis prediction for each patient \( u_i \) is to predict the diagnosed diseases in \( Q_i = \{d_{i,j}, T_{i,j}\}_{j=N_i}^{N_i+1} \) based on \( S_i = \{d_{i,j}, T_{i,j}\}_{j=1}^{N_i} \), where \( N_S \) denotes a small number of diagnosed diseases in the previous clinical visits of \( u_i \). For unified training, we normalize the diagnosis time \( T_{i,j} \) into \([0,1] \) via \((T_{i,j}−T_{i,0})/(T_{\text{max}}−T_{i,0})\).

3.2 Overall Framework
We summarize our MetaCare framework as a two-step process (shown in the left of Figure 2) as follows:

(1) We denote the embedding for a disease as \( d_j \in \mathbb{R}^D \), where \( D \) is the embedding dimension. To model the accumulative effects of disease progression over time, we propose to explicitly stack the previous diagnosed diseases. In other words, when predicting diagnosis in \( \tau_{i,j+1} \), we leverage \( \tau_{i,j}[1] = \{d_{i,j,x}, T_{i,j,x}\}_{x=1}^{N_i} \) as inputs. In this way, we can obtain a latent vector \( z \) in an iterative way, where \( z \) is learned via encoder \( h_\theta \) (Section 3.3.1). The generation process can be formulized as follows:

\[
p(i, N; \tau_{i,[1:N−1]} = \int p(z_i) \prod_{j=1}^{N_i−1} p(\tau_{i,j+1}|\tau_{i,[1:j]}, z_i) \, dz_i
\]

(1)

Since the true posterior \( p \) in Eq. 14 is intractable, we propose to leverage continuous normalizing flow (CNF) to infer it (Section 3.3.2). Through a series of invertible mappings that transform an initial latent variable as \( z_1(\Psi) \) to a more complicated one as \( z_i(\Psi) \), we can alleviate the inference gap and achieve better performance. The final variational posterior is defined as \( q_\phi(z_i(\Psi)|\tau_i) \). We have the evidence lower-bound (ELBO) objective as follows:

\[
\text{arg} \max_{\theta, \phi} \sum_{i} \text{ELBO} \left( q_\phi(z_i(\Psi)|\tau_i) \log p(z_i(\Psi), \tau_{i,[j+1]}(\Psi)) \right)
\]

(2)

(2) To predict diagnosis in the query seq \( Q_i \), we leverage the variable \( z_i(\Psi) \) with \( S_i \) by replacing the Matching part above as follows:

\[
\text{E}_{q_\phi(z_i(\Psi)|\tau_i)} \sum_{j} \prod_{j} p(\tau_{i,j+1}|\tau_{i,[1:j]}, z_i(\Psi)) - \log q_\phi(z_i(0)||\tau_i) p(z_i(0))\]

(3)

where the conditional likelihood \( p(\tau_{i,j+1}|\tau_{i,[1:j]}, z_i(\Psi)) \) is learned by decoder \( g_\nu \) (Section 3.4).

3.3 Encoder
3.3.1 Embedding Layer. Given \( \tau_i \) in the training set and \( S_i \) in the testing set, our encoder \( h_\theta \) tries to generate the variational approximations \( q(z_i|\tau_i) \) and \( q(z_i|S_i) \) respectively. To capture the temporal relations among sequential diagnoses, especially the accumulative effects of disease progression over time, we propose to explicitly model the diagnosis time of each disease. In particular, we first concatenate the previous diagnosed diseases before
the diagnosis time of $\tau_{i,1:j}$ and the predictive ones in $\tau_{i,j+1}$ as $E^x_{i,1:j+1} = [u_i, d_1, T_{i,1}, \ldots, d_j, T_{i,j+1}]$, where the $u_i$ is the patient embedding. Then, we apply a multilayer perceptron network (MLP) to obtain the representation of $j$-th disease $E^x_{i,j} = MLP(E^x_{i,j-1})$. In this way, the diagnosed diseases at the early period are explicitly encoded into the $i$-th patient’s current physician state $E^x_{i,j}$. Then, we further apply an attention mechanism across $E^x_{i,1}, E^x_{i,2}, \ldots, E^x_{i,N-1}$, and obtain the representation of each sequence as $s_i = \text{softmax}(E^x_iW_s)E_i$, where $W_s$ is learnable parameters.

### 3.3.2 Inference with CNF

With the sequence representation $s_i$ learned, it is widely adopted to leverage the reparameterization trick [12, 15] to express the random variable, i.e., $z_i ~ \mathcal{N}(\mu_i, \text{diag}(\sigma_i^2))$ as follows:

$$ [\mu_i, \sigma_i] = \text{Encoder}(s_i; h_0),$$  

$$z_i = \mu_i + \epsilon \odot \sigma_i, \quad \epsilon \sim \mathcal{N}(0, I). \quad (4)$$

Although it is possible to use a simple Gaussian prior over the real posterior distribution $q_{\phi}(z_i|\tau_i)$, a restricted prior as a typical Gaussian in Eq. 4 tends to limit the model performance [5, 40, 45]. Existing studies have shown that, by employing richer and more complicated distributions, we can alleviate the non-negligible inference gaps between the true posterior $p$ and the approximate posterior $q_{\phi}$. However, the inference of such expressive probability distributions is non-trivial.

To this end, we propose to leverage a continuous normalizing flow (CNF), which provides a general and extensible framework for modelling highly complex distributions [3, 4, 11, 25]. In particular, we simplify the computation of the change in $z_i$ and its log densities to transform $q(z_i|\tau_i)$ in a continuous way. Given a variable $z_i(0)$ with known probability distribution $p(z_i(0))$ (e.g., Gaussian in Eq. 4) and differential function $\beta_{\xi}$ that is uniformly Lipschitz continuous in both $z_i$ and step $\psi$, we have:

$$\frac{dz_i(\psi)}{d\psi} = \beta_{\xi}(z_i(\psi), \psi). \quad (5)$$

which describes a continuous-in-time transformation of $z_i(\psi)$. With the theorem of instantaneous change of variables [4], the change in log densities $\log q(z_i(\psi)|\tau)$ also follows a differential equation:

$$\frac{d\log q(z_i(\psi)|\tau)}{d\psi} = -\text{Tr}\left(\frac{\partial \beta_{\xi}(z_i(\psi), \psi)}{\partial z_i(\psi)}\right), \quad (6)$$

where Tr denotes the trace operation and can replace the intensive determinant computation in normalizing flows [46].

Then, the latent variables $z_i(\Psi)$ after step $\Psi$ and its log densities $\log q(z_i(\Psi)|\tau)$ can be computed as:

$$z_i(\Psi) = z_i(0) + \int_0^\Psi \beta_{\xi}(z_i(\psi), \psi) d\psi, \quad (7)$$

$$\log q_{\phi}(z_i(\Psi)|\tau) = \log q_{\phi}(z_i(0)|\tau) - \int_0^\Psi \text{Tr}\left(\frac{\partial \beta_{\xi}(z_i(\psi), \psi)}{\partial z_i(\psi)}\right) d\psi,$$

where $\Psi$ can be arbitrarily set for more transformations and we empirically set $\Psi$ as 1 following [45].

With log $q_{\phi}(z_i(\Psi)|\tau)$ in Eq. 7, the matching part in Eq. 2 can be calculated based on CNF as follows:

$$L_m = \mathbb{E}_{q_{\phi}(z_i|\tau_i)} \log \frac{q_{\phi}(z_i(0)|\tau_i)}{p(z_i(0))} = \mathbb{E}_{q_{\phi}(z_i|\tau_i)} \log p(z_i(\Psi)) \quad (8)$$

$- q_{\phi}(z_i(0)|\tau_i) + \int_0^\Psi \text{Tr}\left(\frac{\partial \beta_{\xi}(z_i(\psi), \psi)}{\partial z_i(\psi)}\right) d\psi.$

### 3.4 Decoder

Due to the inverse property of continuous normalizing flow [38], there exists condition of consistency between $q_{\phi}(z_i(0)|\tau_i)$ and

---

**Figure 2: The overall framework of both MetaCare and MetaCare++**.
where \( \tau_{i,j+1} = (d_{i,j+1}, T_{i,j+1}) \) includes the diagnosis \( d_{i,j+1} \) and the exact diagnosis time \( T_{i,j+1} \). In particular, we first obtain the temporal context representation \( \Gamma_{i,j} \in \mathbb{R}^{D \times 1} \) as follows:

\[
\Gamma_{i,j} = \text{MLP}((\tau_{i,j}, z_i(\Psi))),
\]

where \( \tau_{i,j} \) is the input and \( z_i(\Psi) \) is the latent variable in Eq. 7.

**Diagnosis prediction layer** aims to predict the diagnosed diseases of each patient. Given the latent variable \( z_i(\Psi) \) in Eq. 7, the probability of the predictive disease \( d_{i,j+1} \) is calculated as follows:

\[
p(d_{i,j+1}|\Gamma_{i,j}) = \text{softmax}(\Gamma_{i,j}).
\]

**Time prediction layer** aims to predict the exact diagnosis time of the next visit. Rather than directly predicting time as a regression problem, we propose to learn a condition intensity function \( \lambda(i,t) \), which can calculate the accumulative influence among the past diagnosis and reflect the evolutionary process of the intensity function with time as follows

\[
\lambda_i(t) = \exp \left( \Gamma_{i,j} \cdot \sigma' + \alpha (t - T_{i,j}) + \lambda_0 \right),
\]

where \( \sigma' \in \mathbb{R}^{D \times 1} \) and \( \lambda_0 \) are scalars that denote the basic intensity of the next diagnosis. The first term \( \Gamma_{i,j} \cdot \sigma' \) calculates the accumulative effect among the past diseases. The second term \( \alpha (t - T_{i,j}) \) denotes the evolutionary process of the intensity function with time.

In the training phase, the probability that the next diagnosed disease would occur at time \( T_{i,j+1} \) can be calculated as:

\[
p(\hat{T}_{i,j+1} = T_{i,j+1}|\Gamma_{i,j}) = f_i(\hat{T}_{i,j+1} - T_{i,j}) = f_i(\Delta T_{i,j+1}),
\]

where \( \Delta t = \lambda_i(t) \exp \left( \frac{-f_i'}{\lambda_i(t)} \right) \). In the test phase, the time \( \hat{T}_{i,j+1} \) of the next diagnosis can be predicted as the expectation

\[
\hat{T}_{i,j+1} = \int_{T_{i,j}}^{\infty} t \cdot f_i(t) dt.
\]

### 3.5 Overall Objectives

Finally, the objective function of the proposed MetaCare is given as follows:

\[
\min L_{\text{pred}} = - \sum_{i=1}^{|\mathcal{S}|} \sum_{j=1}^{N_{2i}-1} \log p(d_{i,j+1}|\Gamma_{i,j})
\]

\[
+ \log p(\hat{T}_{i,j+1} = T_{i,j+1}|\Gamma_{i,j}) - L_m,
\]

where \( p(d_{i,j+1}|\Gamma_{i,j}) \) is calculated in Eq. 11, \( p(\hat{T}_{i,j+1} = T_{i,j+1}|\Gamma_{i,j}) \) is calculated in Eq. 13, and the matching loss \( L_m \) is calculated in Eq. 8. \( d_{i,j} \) and \( T_{i,j} \) are the ground-truth labels of the \( (j+1) \)-th diagnosis and its diagnosis time.

### 4 METACARE WITH HIERARCHICAL SUBTypING (METACARE++)

Our proposed MetaCare framework essentially captures the temporal relations among sequential diagnoses, which is helpful to model infrequent patients. However, the current modeling of diseases fails to capture either the hierarchical relations or syndromic relations among diseases. Moreover, we find that generally sharing parameters for all patients’ sequential diagnoses can be problematic, since there exist specific causes of diseases to different patients. Without tailoring personalized parameters for individual patients, the model with globally shared knowledge can even backfire personalized diagnosis prediction.

In light of this, we propose a diagnosis-enhanced disease representation learning method to capture both hierarchical and syndromic relations among diseases, and further propose a personalized decoder to tailor the meta-model’s parameters via patient subtyping (shown in the right of Figure 2). In this way, we can jointly model infrequent patients and rare diseases for personalized diagnosis prediction. We name our framework of MetaCare with hierarchical subtyping as MetaCare++.

### 4.1 Diagnosis-enhanced Disease Representation

#### 4.1.1 Capturing Hierarchical Relations based on ICD-9 Hierarchy.

ICD-9 is an official system of assigning medical codes to diseases [34]. It hierarchy-classifies medical codes into different types of diseases according to the body systems in L levels. This forms a tree structure where each disease has only one direct parent. To ensure the parent disease summarize the common properties of child diseases and child diseases inherit important properties from parents, we recursively concatenate the embedding of each sub-disease to their parent from the leaf level of hierarchy as follows:

\[
d_j = E_j^1 \oplus E_j^2 \oplus \cdots \oplus E_j^L \in \mathbb{R}^{D \times 1},
\]

where \( \oplus \) denotes the concatenation operation and \( E_j^l \) denotes the disease representation in the \( l \)-th level.

#### 4.1.2 Capturing Syndromic Relations via Hierarchy-aware Disease Subtyping.

Besides the hierarchical relations, it is important to mine the syndromic relations among diseases as motivated in Section 1. Although the syndromic relations can be obtained via actual diagnosis data (i.e., the patient-disease matrix \( \Upsilon \)), it is non-trivial to make disease embeddings reflect the syndromic relations and preserve the hierarchy inherited from ICD-9. For example, as shown in Figure 1, with the syndromic relation between 401.0 and 790.21, directly forcing 790.21 to be close to 401.0 may break the parent-child relation between 790.2 and 790.21.

To properly capture the syndromic and hierarchical relations among diseases, we propose to reorganize the embedding space based on the ICD-9 hierarchy. Specifically, we first capture the syndromic relations among diseases via disease subtyping, which clusters diseases into different groups according to the patient-disease matrix. Then, we propose to calculate the center of each disease subtype by considering the hierarchical relations, so as to reorganize disease embeddings via subtype-aware regularization loss. The detailed processes are listed as follows:
Capturing syndromic relations via disease subtyping: Suppose the diseases can be divided into \( K \) subtypes as \( \mathcal{G} = \{ \mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_K \} \), where \( \mathcal{G}_k (1 \leq k \leq K) \) denotes a disease subtype and \( K \) is a hyper-parameter. Through parent-disease matrix \( \mathbf{Y} \), we can obtain the concurrent frequency of two diseases. Specifically, we first multiply \( \mathbf{Y}^\top \mathbf{Y} \) to calculate the disease-disease relations, where the value in the matrix represent the concurrent times of two diseases. Then, we divide the concurrent times by the total concurrent times of each disease for the concurrent frequency. For example, as shown in Figure 1, both Lisa and Jack have been diagnosed with disease 272.0 and disease 401.0. In this case, the concurrent times between 272.0 and 401.0 is 2. By leveraging the concurrent frequency as the similarity, we can simply perform K-Means [17] for disease subtyping.

Preserving hierarchical relations based via an accumulative regularization: With the disease subtypes \( \mathcal{G}_k \), it is straightforward to reorganize the embedding space by regularizing the diseases inside one subtype to be closed to the center of the subtype. However, without considering that parent diseases can summarize the properties of their children, the hierarchical relations may be destroyed after the mining of syndromic relations. For example, as shown in Figure 1, the syndromic relations between 272.0 and 401.0 should not only make themselves close to each other but also ensure their parents (i.e., 272 and 401) to be close to each other.

In light of this, we propose to disentangle the disease embeddings to \( \mathbf{E}_i^l \) in Eq. 15 and perform an accumulate regularization from the sub-diseases to the parent-diseases. In this way, we can regularize the disease embeddings to be closer to the center of the disease subtypes at each level, to make the embedding space after capturing the syndromic relations consistent with the ICD-9 hierarchy. The proposed accumulative regularization loss is given as follows:

\[
\mathcal{L}_{d-subtyping} = \sum_{k=1}^{K} \sum_{d_i \in \mathcal{G}_k} \sum_{l=1}^{l_i} \| \mathbf{E}_i^l - \frac{1}{m_k} \sum_{d_j \in \mathcal{G}_k} \mathbf{E}_j^l \|^2, \tag{16}
\]

where \( \| \cdot \|^2 \) measure the Euclidean distance and \( \mathbf{E}_i^l \) is the \( l \)-th disease’s embedding at level \( l \), \( l_i \) denotes \( d_i \)’s actual level in ICD-9 and \( m_k \) is the number of diseases that belong to subtype \( \mathcal{G}_k \).

4.2 Personalized Decoder

The proposed MetaCare++ in Section 3 captures the temporal relations among sequential diagnosis, and the diagnosis-enhanced disease representation learning in Section 4.1 captures the complicated relations among syndromic diseases. Both of them are leverage globally shared meta-learning parameters for diagnosis prediction. However, since the leading causes of one specific disease for different patients vary a lot [21, 41], the general meta-learner fails to handle the patient personalization and can backfire personalized diagnosis prediction. Since the personalized parameters for each patient will involve a large memory cost and suffer from the data sparsity problem, we propose to perform patient subtyping to personalize meta-learner for a certain group of patients.

Personalizing the meta-learner via patient subtyping: To perform an accurate patient subtyping, we propose to enhance the representation of patients’ diagnosis sequences by aggregating the disease embeddings. With such a disease-enhanced patient embedding \( \mathbf{u}_i \), the patient embeddings can inherit both the hierarchical and syndromic relations among diseases. Since \( \mathbf{u}_i \) is the linear combination of the diagnosed disease, it can also be disentangled into \( L \) level representation similar to the disease representation in Eq. 15. The formulation of \( \mathbf{u}_i \) is given as follows:

\[
\mathbf{u}_i = \frac{\sum_{l \in \mathcal{G}_k} v_l \mathbf{d}_l}{\sum_{l \in \mathcal{G}_k} v_l} = \mathbf{E}^{u_1}_i \oplus \mathbf{E}^{u_2}_i \oplus \cdots \oplus \mathbf{E}^{u_L}_i, \tag{17}
\]

where \( v_l \in \mathcal{Y} \) is from patient-disease matrix and \( \mathbf{E}^{u_l}_i \) represents the \( l \)-th level feature of patient \( \mathbf{u}_i \).

Then, since \( \mathbf{u}_i \) inherits the hierarchical property of disease by aggregation, we propose to perform hierarchical subtyping, where the patient subtypes are denoted as \( \mathcal{H} = \{ \mathcal{H}_1^1, \mathcal{H}_2^1, \ldots, \mathcal{H}_C^1, \ldots, \mathcal{H}_C^L \} \). \( C \) is the number of child subtypes of each level and \( L \) is the number of level that is consistent with the ICD-9 hierarchy. The corresponding subtype representation is initialized as \( \mathcal{H} = \{ \mathcal{H}_1^1, \mathcal{H}_2^1, \ldots, \mathcal{H}_C^1, \ldots, \mathcal{H}_C^L \} \). To measure the importance of different subtypes in \( \mathcal{H} \) for \( \mathbf{u}_i \), we calculate the attention score \( \alpha_{i,j}^l \) between patient \( \mathbf{u}_i \) and the subtype \( \mathcal{H}_j^l \) at the \( l \)-th level as:

\[
\alpha_{i,j}^l = \frac{\| \mathbf{E}^{u_l}_i - \mathbf{H}_j^l \|^2}{\sum_{f} \| \mathbf{E}^{u_l}_i - \mathbf{H}_f^l \|^2}. \tag{18}
\]

Therefore, the enhanced patient representation \( \mathbf{P}_i \) can be obtained by fusing the knowledge from both disease-enhanced patient embedding and patient’s subtype embedding as follows:

\[
\mathbf{P}_i = \text{MLP}(\mathbf{u}_i + \sum_{l=1}^{L} \sum_{j=1}^{K} \alpha_{i,j}^l \mathbf{H}_j^l). \tag{19}
\]
Similar to the regularized loss for reorganizing disease embeddings in Eq. 16, we propose to update subtype embeddings of patients via minimizing the distance between average patient embeddings and the center $\mathbf{H}_j^l$ at different levels $l$ as follows:

$$L_{p\text{-subtyping}} = \sum_{l=1}^{L} \sum_{j=1}^{L} \sum_{i \in \mathcal{H}_j^l} \left\| \mathbf{H}_j^l - \mathbf{P}_i \circ \mathbf{g}_v \right\|_2^2,$$  \hspace{1cm} (20)

where $n_j^l$ is the number of patients in subtype $\mathcal{H}_j^l$.

Finally, we learn a personalized decoder $g_v$ to adapt only the personalized parameters from general decoder $g_v$, as follows:

$$g_v = \mathbf{P}_i \circ g_v,$$ \hspace{1cm} (21)

where $\circ$ denotes the element-wise multiplication and $g_v$ is learned via Eq. 10-Eq. 13.

**Connection between disease subtyping and patient subtyping:** Performing disease subtyping and patient subtyping together in a unified framework would allow them to mutually reinforce each other. The hierarchical and syndromic relations among diseases can be captured after the disease subtyping, which can be helpful to represent patients and personalize the meta-learner via patient subtyping. Moreover, the personalized meta-learner can also be helpful to update the disease embeddings, since the disease subtyping is in an unsupervised fashion and based on the embeddings learned from the meta-learning framework.

### 4.3 Overall Objectives

The final objective function of MetaCare++ is given as follows:

$$\min L_{pred} + \omega (L_{d\text{-subtyping}} + L_{p\text{-subtyping}}),$$ \hspace{1cm} (22)

where $L_{pred}$ is calculated in in Eq. 14 for both diagnosis and diagnosis times prediction tasks. $\omega$ is a hyperparameter to control the weight for the subtyping of diseases (i.e., $L_{d\text{-subtyping}}$ in Eq. 16) and patients($L_{p\text{-subtyping}}$ in Eq. 20).

### 5 EXPERIMENT

In this section, we evaluate our proposed MetaCare and MetaCare++ frameworks focusing on the following four research questions:

- **RQ1:** How do MetaCare and MetaCare++ perform in comparison to state-of-the-art diagnosis prediction methods?
- **RQ2:** What are the effects of different model components?
- **RQ3:** How do the hyperparameters affect the prediction performance and how to choose optimal values?
- **RQ4:** How does MetaCare++ improve the modeling of infrequent patients and rare diseases?

### 5.1 Experimental Setup

#### 5.1.1 Datasets and Evaluation Protocols

We use two real-world EHR datasets to verify the effectiveness of compared methods, i.e., MIMIC-III [13] and eICU [28]. Both datasets are fully anonymized and carefully sanitized before our access. We chose patients who made at least two visits for both datasets. Both datasets have $L = 4$ levels in ICD-9 hierarchy. The statistics are summarized in Table 1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MIMIC-III</th>
<th>eICU</th>
</tr>
</thead>
<tbody>
<tr>
<td># of patients</td>
<td>7,499</td>
<td>11,707</td>
</tr>
<tr>
<td># of visits</td>
<td>19,911</td>
<td>25,661</td>
</tr>
<tr>
<td>Avg. visits per patient</td>
<td>2.66</td>
<td>2.19</td>
</tr>
<tr>
<td># of unique ICD9 codes</td>
<td>6,984</td>
<td>941</td>
</tr>
<tr>
<td>Avg. # of diagnosis codes per visit</td>
<td>8.78</td>
<td>4.82</td>
</tr>
<tr>
<td>Max # of diagnosis codes per visit</td>
<td>39</td>
<td>57</td>
</tr>
</tbody>
</table>

#### 5.1.2 Methods for comparison

We compare both MetaCare and MetaCare++ with the following baselines from two perspectives: (1) existing diagnosis prediction methods (RETAIN [6], Dipole [22], Timeline [2], GRAM [7], KAME [23], MHM [31], TAdaNet [36], and CGL [20]); (2) existing meta-learning methods for cold-start users (MeLU [16], MAMO [8], and TaNP [18]):

- **RETAIN** [6] is a diagnosis prediction model that leverages GRUs and attention mechanisms to calculate the contribution scores of all the appeared diagnosis codes.
- **Dipole** [22] uses bidirectional RNNs and attention mechanisms to predict patient visit information.
- **Timeline** [2] devise a time-aware disease progression function to predict clinical events from past visits.
- **GRAM** [7] uses a medical knowledge graph to learn the medical code representations and predict the future visit information with recurrent neural networks.
- **KAME** [23] is a model for predicting patients’ future health information based on knowledge attention mechanism.
- **MHM** [31] models multi-modal clinical data-based hierarchical multi-label model, which integrates discrete medical codes, structural information and time series data into the same framework for the diagnosis prediction task.
- **TAdaNet** [36] proposes a task-adaptive network that makes use of a domain-knowledge graph to enrich data representations and provide task-specific customization for rare disease detection.
- **CGL** [20] design a collaborative graph learning model to explore patient-disease interactions and medical domain knowledge.
- **MeLU** [16] handles cold-start user modeling by applying the framework of MAML [10]. Based on the learned parameter initialization, MeLU makes recommendations for cold-start users via a few steps of gradient updates.
- **MAMO** [8] designs the task-specific and feature-specific memory matrices for user cold-start and item cold-start problems based on a memory-augmented framework of MAML.
- **TaNP** [18] maps the observed interactions of each user to a predictive distribution and learns a task-specific meta-learning framework for user cold-start recommendation.

#### 5.1.3 Implementation Details

We implement both MetaCare and MetaCare++ with Pytorch\(^1\), which will be fully released upon the acceptance of this work. Implementations of the compared baselines are either from open-source projects or the original authors.

\(^1\)https://pytorch.org/
We compare the cold-start diagnosis prediction results of the proposed MetaCare and MetaCare++ framework to those of the baseline models. Table 2 shows the Recall @K, NDCG@K, and MAE scores on MIMIC-III and eICU datasets with K=[5,10]. We have the following observations.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall@5</th>
<th>NDCG@5</th>
<th>Recall@10</th>
<th>NDCG@10</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>RATAIN</td>
<td>0.0919</td>
<td>0.2899</td>
<td>0.1388</td>
<td>0.2663</td>
<td>0.1626</td>
</tr>
<tr>
<td>Dipole</td>
<td>0.0724</td>
<td>0.2579</td>
<td>0.1257</td>
<td>0.2587</td>
<td>0.1943</td>
</tr>
<tr>
<td>Timeline</td>
<td>0.0853</td>
<td>0.2636</td>
<td>0.1282</td>
<td>0.2622</td>
<td>0.1710</td>
</tr>
<tr>
<td>GRAM</td>
<td>0.1045</td>
<td>0.3265</td>
<td>0.1702</td>
<td>0.2845</td>
<td>0.1577</td>
</tr>
<tr>
<td>KAME</td>
<td>0.1033</td>
<td>0.3218</td>
<td>0.1696</td>
<td>0.2793</td>
<td>0.1609</td>
</tr>
<tr>
<td>MHM</td>
<td>0.1056</td>
<td>0.3289</td>
<td>0.1756</td>
<td>0.3054</td>
<td>0.1504</td>
</tr>
<tr>
<td>TAdaNet</td>
<td>0.1094</td>
<td>0.3316</td>
<td>0.1792</td>
<td>0.3130</td>
<td>0.1493</td>
</tr>
<tr>
<td>CGL</td>
<td>0.1174</td>
<td>0.3438</td>
<td>0.1793</td>
<td>0.3160</td>
<td>0.1319</td>
</tr>
<tr>
<td>MeLU</td>
<td>0.0945</td>
<td>0.3208</td>
<td>0.1507</td>
<td>0.2887</td>
<td>0.1443</td>
</tr>
<tr>
<td>MAMO</td>
<td>0.1004</td>
<td>0.3354</td>
<td>0.1547</td>
<td>0.3147</td>
<td>0.1408</td>
</tr>
<tr>
<td>TaNP</td>
<td>0.1012</td>
<td>0.3395</td>
<td>0.1587</td>
<td>0.3151</td>
<td>0.1322</td>
</tr>
<tr>
<td>MetaCare</td>
<td>0.1195</td>
<td>0.3478</td>
<td>0.1831</td>
<td>0.3214</td>
<td>0.1293</td>
</tr>
<tr>
<td>MetaCare++</td>
<td>0.1296*</td>
<td>0.3725*</td>
<td>0.1920*</td>
<td>0.3486*</td>
<td>0.1107*</td>
</tr>
</tbody>
</table>

In general, both MetaCare and MetaCare++ outperform all 11 baselines across all evaluation metrics on both datasets. This answers RQ1, showing that our proposed specific clinical meta-learning framework is capable of precise diagnosis prediction. Moreover, the ranking of many baselines is fluctuating across datasets as we see the second-best performance scattered among different models like CGL and TaNP. Compared with the second-best performance, the performance gains of MetaCare++ in terms of Recall and NDCG range from reasonably large (5.07% achieved with Recall@10 on eICU) to significant large (10.39% achieved with Recall@5 on MIMIC-III). The performance gains of MetaCare++ in terms of MAE range from 11.81% on eICU to 16.07% on MIMIC-III.

In particular, by considering latent temporal relations among patient visits and collaboratively learning representations of patients and diseases, CGL performs better than TaNP in many cases. Compared with CGL, MetaCare++ not only takes the data sparsity problem into consideration but also explicitly models the accumulated effects of disease progression. Therefore, MetaCare++ outperforms CGL by up to 10.39% in Recall@5 on MIMIC-III on the diagnosis prediction task and up to 16.07% on MIMIC-III on the diagnosis time prediction task.

Since CGL can suffer from the data sparsity problem (e.g., as shown in Table 1, the average number of visits of each patient on the eICU dataset is 2.19), TaNP can sometimes achieve better performance since it can learn parameters by a few steps of gradient updates. The main differences between MetaCare++ and TaNP reside in properly capturing the temporal relations among clinical visits and complicated relations among syndromic diseases. Specifically, MetaCare++ can outperform TaNP by up to 28.06% in Recall@5 on MIMIC-III on the diagnosis time prediction task and 12.84% on eICU on the diagnosis time prediction task.

5.3 Model Ablation (RQ2)

To better understand our proposed techniques, i.e., continuous normalizing flow (CNF), clinical meta-learner jointly regarding
diagnoses and diagnosis times, diagnosis-enhanced disease representation learning (Diag-Disease), and personalized decoder, we study MetaCare++ as follows:

- Meta is the variational autoencoder based meta-learning model for diagnosis prediction, which does not involve the task of diagnosis time prediction;
- Meta + CNF is the meta-learning framework based on continuous normalizing flow, which employs complex distributions to alleviate the inference gaps;
- MetaCare is our proposed clinical meta-learning framework, which captures temporal relations among clinical visits of individual patients regarding both diagnoses and times;
- MetaCare + Diag-Disease is the MetaCare model with diagnosis-enhanced hierarchical disease representation, which capture both hierarchical and syndromic relations among diseases based on the domain knowledge and the actual diagnosis data;
- MetaCare++ integrates MetaCare + Diag-Disease with a personalized decoder to bridge the modeling of both infrequent patients and rare diseases.

From Table 3, we have the following observations:

The performance gains of Meta + CNF over Meta on two datasets fluctuate, ranging from 1.61% (achieved in Recall@5 on eICU) to 12.95% (achieved in Recall@5 on MIMIC-III). Similarly, the corresponding performance gains of MetaCare over Meta + CNF ranges from 1.92% (achieved in Recall@5 on eICU) to 12.31% (achieved in Recall@5 on MIMIC-III). These results show that both CNF and temporal relations among clinical visits can bring enhancement to the generic meta-learning framework for diagnosis prediction.

Furthermore, the performance gains of MetaCare + Diag-Disease over MetaCare ranges from 2.88% (achieved in NDCG@5 on eICU) to 9.90% (achieved in MAE on MIMIC-III). The result shows that: (1) the explicitly modeling hierarchical and syndromic relations among diseases can further improve the performance of clinical meta-learner, where MetaCare does not consider such domain knowledge from ICD-9 and actual diagnosis; (2) on the datasets (e.g., MIMIC-III) that have a larger number of diseases, the improvements of diagnosis-enhanced representation learning are more significant by properly arranging disease embedding according to the context and structure information in the taxonomy.

Compared with MetaCare + Diag-Disease, MetaCare++ leads to performance gains ranging from 1.18% (achieved in NDCG@5 on eICU) to 4.98% (achieved in MAE on MIMIC-III). Even though MetaCare + Diag-Disease has already integrated the complicated relations among syndromic diseases, MetaCare++ can still improve the performance by performing hierarchical patient subtyping and tailoring personalized parameters for each patient.

### 5.4 Effect of Hyperparameters (RQ3)

Our proposed MetaCare++ framework mainly introduces three hyperparameters, i.e., $K$, $C$, and $\omega$.

From Figure 3, we have the following observations: (1) $K$ is used for disease subtyping, where we found that the optimal $K$ is about 10. (2) $C$ is used for hierarchical patient subtyping, where we found that the optimal $C$ is about 4. The rules for selecting $K$ and $C$ could be the rule-of-thumb in practice across the used datasets. (3) $\omega$ controls the weight of both disease subtyping and patient subtyping, which aims to enforce the disease embeddings to be close to the weighted center of nodes in the taxonomy and as do patient embeddings. Too small $\omega$ will cause the tag embeddings likely to be spread out, while too large $\omega$ will likely cause the model to overfit. The optimal $\omega$ value on MIMIC-III is about 0.1. Note that, MetaCare++ is reasonably sensitive to $\omega$. In the range of $[0.1, 1]$, the optimal $\omega$ can be obtained by slight tuning.

![Figure 3: Performance regarding Recall@5 of the best baseline, MetaCare, and MetaCare++ on the MIMIC-III dataset.](image-url)
5.5 Case Studies (RQ4)
To provide more insights into the advantages of MetaCare++ in modeling infrequent patients and rare diseases, we provide the visualized embedding space of diseases and demonstrate examples of infrequent patient cases as follows.

As shown in Figure 4, we visualized the first 16 dimension embedding vectors learned by the proposed MetaCare and MetaCare++ on MIMIC-III. MetaCare and MetaCare++ use the same color spectrum to represent different disease categories according to ICD-9 disease hierarchy. In MetaCare, the model mainly captures the temporal relations among clinical visits, whereas MetaCare++ additionally captures the hierarchical and syndromic relation among diseases together with the temporal relations. From Figure 4, it is hard to find regularity in the distribution of diseases from different categories in the embedding space learned by MetaCare. However, the diseases from different categories are well separated in the embedding spaces of MetaCare++. For example, Diseases Of The Circulatory System (390-459) in red is close to Diseases Of The Musculoskeletal System And Connective Tissue (710-739) in green while is far away from Mental Disorders (290-319) in blue. Moreover, the embedding spaces do include different syndromic relations of diseases, such as part of 390-459 is concurrent with 710-739 while part of 390-459 is concurrent with 270-279.

To provide more insights, we further demonstrate three examples of infrequent patients (as shown in Table 4), the diagnosis predictions are made by MetaCare++. Although the three patients have suffered from the same high blood pressure disease that belongs to the cardiovascular system (i.e., 401.0), they can belong to different patient subtypes. Since Linda suffers from 272.0 that belongs to disorders of lipid metabolism, the concurrent 272.0 and 401.0 leads to a risk of Metabolic syndrome rather than pure Cardiovascular Syndrome as Lisa and Jack. Based on the above subtypes, we can make personalized diagnosis predictions for Linda as 790.5 that belong to the abnormal examination of blood, whereas the prediction for Lisa and Jack still belong to the cardiovascular system. Note that, the exact subtype labels we create here are not perfectly accurate due to the implicit nature of patient subtyping. However, they nonetheless provide valuable insight into the meaningful representative patients directly extracted from the implicit diagnosis data in an unsupervised fashion, which is helpful to personalized meta-learner for different parameters.

Table 4: Examples of infrequent patients modeled by the proposed MetaCare++ and the corresponding diagnosis prediction in MIMIC-III.

<table>
<thead>
<tr>
<th>Visit</th>
<th>Subtype</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lisa 431.401.0;...</td>
<td>Cardiovascular Syndrome</td>
<td>414.0;...</td>
</tr>
<tr>
<td>Jack 435.9; 401.0;...</td>
<td>Cardiovascular Syndrome</td>
<td>427.1;...</td>
</tr>
<tr>
<td>Linda 272.0; 401.0;...</td>
<td>Metabolic syndrome</td>
<td>790.5;...</td>
</tr>
</tbody>
</table>

6 CONCLUSION
In this paper, we propose to make diagnosis predictions for both infrequent patients and rare diseases regarding both diagnoses and times. Specifically, we propose a novel specialized clinical meta-learner with a hierarchical subtyping strategy (MetaCare++), which captures the temporal relations among patient visits together with the complicated relations among syndromic diseases in a unified framework. Extensive quantitative experiments demonstrate the clear advantages of our MetaCare++ over the state-of-the-art baselines towards the precise diagnosis and the diagnosis time prediction, which is further consolidated with our real case study results.

In the future, it would be interesting to consider the incorporation of explicit patient attributes and disease properties when they are available. Moreover, the inferred subtypes of clinical entities (e.g., patients, diseases, and drugs) from the learned MetaCare++ model can be utilized to support other important healthcare tasks such as risk prediction, patient care, and drug recommendation.

ACKNOWLEDGMENTS
This work was supported in part by the National Natural Science Foundation of China (No.72192823 and No.62172362) and Leading Expert of National “Ten Thousands Talent Program” of Zhejiang Province (No.2021R52001).

REFERENCES