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
xlzheng@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.

SIGIR '22, July 11-15, 2022, Madrid

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**.

KEYWORDS

Diagnosis prediction, Meta-learning, Hierarchical, Subtyping

ACM Reference Format:
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 *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 11-15, 2022, Madrid.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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.,

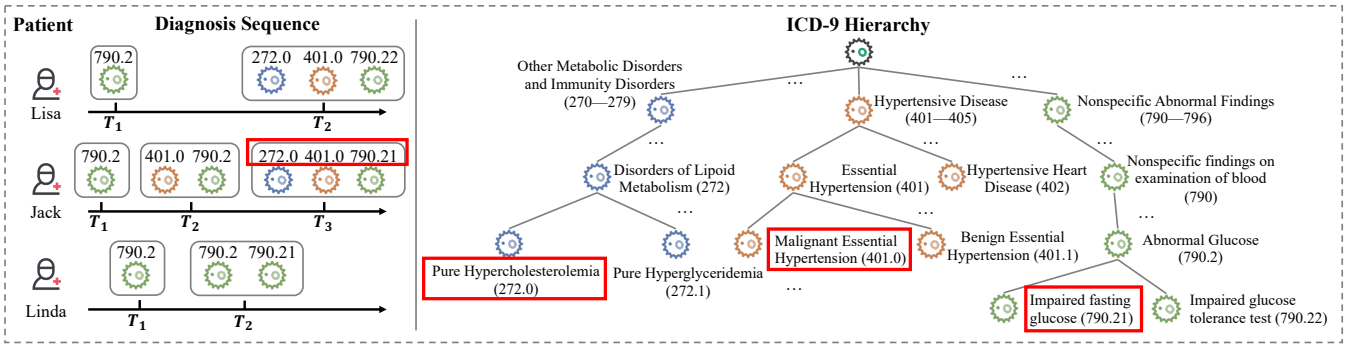


Figure 1: An illustrative example of cold-start diagnosis prediction with patients' sequential diagnoses and ICD-9 hierarchy.

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. TaNP [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 $\mathcal{D}$. For each patient u_i, $\mathcal{D}$ 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 $\mathcal{S}_i$ and a query sequence $\mathcal{Q}_i$ according the diagnosis time ($\tau_i = \mathcal{S}_i \cup \mathcal{Q}_i$). Thus, the task of cold-start diagnosis prediction for each patient u_i is to predict the diagnosed diseases in $\mathcal{Q}_i = \{d_{i,j}, T_{i,j}\}_{j=|\mathcal{S}_i|+1}^{N_i}$ based on $\mathcal{S}_i = \{d_{i,j}, T_{i,j}\}_{j=1}^{N_{\mathcal{S}_i}}$, where $N_{\mathcal{S}_i}$ 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 $(T_{i,j} - T_{i,0})/(T_{max} - T_{i,0})$.

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 $\boldsymbol{d}_j \in \mathbb{R}^D$, 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 $\boldsymbol{\tau}_{i,j+1}$, we leverage $\boldsymbol{\tau}_{i,[1:j]} = \{\boldsymbol{d}_{1,x}, T_{i,x}\}_{x=1}^{j}$ as inputs. In this way, we can obtain a latent vector $\boldsymbol{z}$ in an iterative way, where $\boldsymbol{z}$ is learned via encoder h_θ (Section 3.3.1). The generation process can be formulated as follows:

$$
\begin{aligned}
p\left(\boldsymbol{\tau}_{i,N_i}|\boldsymbol{\tau}_{i,[1:N_i-1]}\right) &= \int p\left(\boldsymbol{z}_i\right) \prod_{j=1}^{N_i-1} p\left(\boldsymbol{\tau}_{i,j+1}|\boldsymbol{\tau}_{i,[1:j]}, \boldsymbol{z}_i\right) d\boldsymbol{z}_i \\
&= \int p\left(\boldsymbol{z}_i\right) \prod_{j=1}^{N_i-1} p\left(\boldsymbol{d}_{i,j+1}, T_{i,j+1}|\boldsymbol{d}_{i,[1:j]}, \boldsymbol{T}_{i,[1:j]}, \boldsymbol{z}_i\right) d\boldsymbol{z}_i.
\end{aligned} \tag{1}
$$

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 $\boldsymbol{z}_i(0)$ to a more complicated one as $\boldsymbol{z}_i(\Psi)$, we can alleviate the inference gap and achieve better performance. The final variational posterior is defined as $q_\phi(\boldsymbol{z}_i(\Psi)|\boldsymbol{\tau}_i)$. We have the evidence lower-bound (ELBO) objective as follows:

$$
\begin{aligned}
&\arg\max_{\theta,\phi} \sum_{j}^{N_i-1} \left[\mathrm{ELBO}\left(q_\phi(\boldsymbol{z}_i(\Psi)|\boldsymbol{\tau}_{i,[1:j]}) \;||\; \log p\left(\boldsymbol{z}_i(\Psi), \boldsymbol{\tau}_{i,j+1}\right)\right)\right] \\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}_i|\boldsymbol{\tau}_i)}[\sum_{j}^{N_i} \underbrace{\log p\left(\boldsymbol{\tau}_{i,j+1}|\boldsymbol{\tau}_{i,[1:j]}, \boldsymbol{z}_i(\Psi)\right)}_{\textit{Reconstruction}} - \underbrace{\log \frac{q_\phi\left(\boldsymbol{z}_i(0)|\boldsymbol{\tau}_i\right)}{p\left(\boldsymbol{z}_i(0)\right)}}_{\textit{Matching}}].
\end{aligned} \tag{2}
$$

(2) To predict diagnosis in the query seq $\mathcal{Q}_i$, we leverage the variable $\boldsymbol{z}_i(\Psi)$ with $\mathcal{S}_i$ by replacing the *Matching* part above as follows:

$$
\mathbb{E}_{q_\phi(\boldsymbol{z}_i|\boldsymbol{\tau}_i)}[\sum_{j}^{N_{Q_i}-1} \log p\left(\boldsymbol{\tau}_{i,j+1}|\boldsymbol{\tau}_{i,[1:j]}, \boldsymbol{z}_i(\Psi)\right) - \log \frac{q_\phi\left(\boldsymbol{z}_i(0)|\boldsymbol{\tau}_i\right)}{q_\phi\left(\boldsymbol{z}_i(0)|\mathcal{S}_i\right)}], \tag{3}
$$

where the conditional likelihood $p\left(\boldsymbol{\tau}_{i,j+1}|\boldsymbol{\tau}_{i,[1:j]}, \boldsymbol{z}_i(\Psi)\right)$ is learned by decoder g_ν (Section 3.4).

3.3 Encoder

3.3.1 Embedding Layer. Given $\boldsymbol{\tau}_i$ in the training set and $\mathcal{S}_i$ in the testing set, our encoder h_θ tries to generate the variational approximations $q(\boldsymbol{z}_i|\boldsymbol{\tau}_i)$ and $q(\boldsymbol{z}_i|\mathcal{S}_i)$ 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

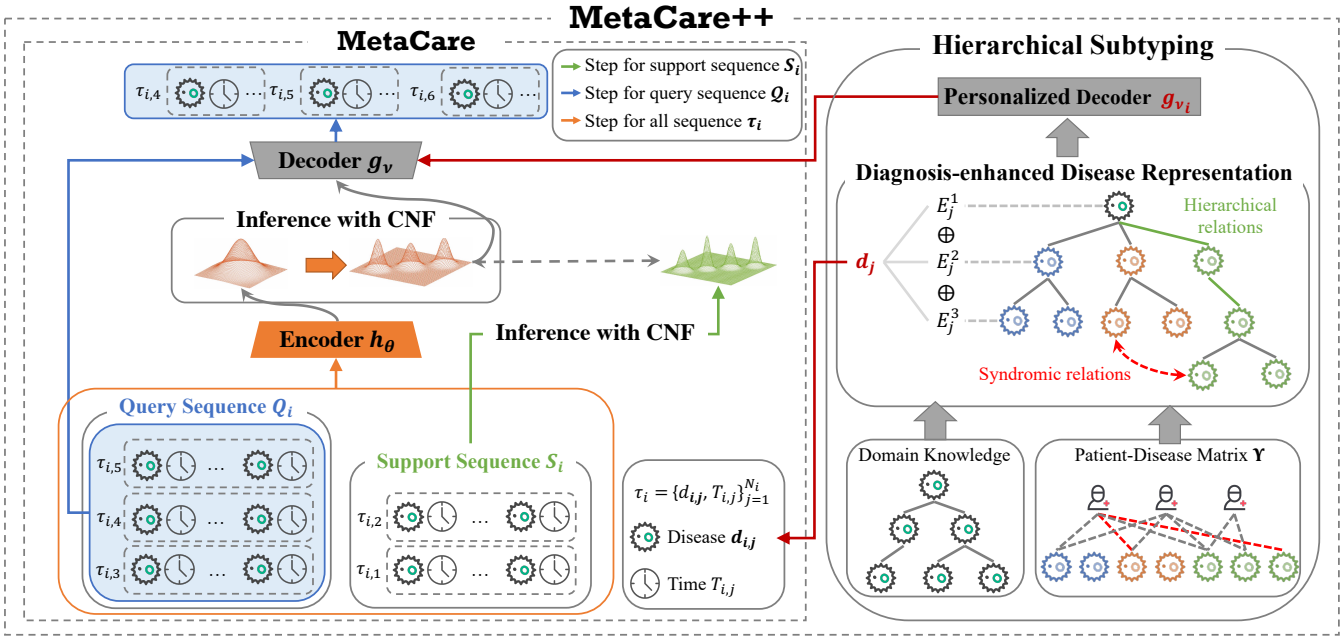


Figure 2: The overall framework of both `MetaCare` and `MetaCare++`.

the diagnosis time of $\boldsymbol{\tau}_{i,[1:j]}$ and the predictive ones in $\boldsymbol{\tau}_{i,j+1}$ as $\boldsymbol{E}^x_{i,[1:j+1]} = [\boldsymbol{u}_i, \boldsymbol{d}_1, T_{i,1}, \ldots, \boldsymbol{d}_{j+1}, T_{i,j+1}]$, where the $\boldsymbol{u}_i$ is the patient embedding. Then, we apply a multilayer perceptron network (MLP) to obtain the representation of j-th disease $\boldsymbol{E}^v_{i,j} = \text{MLP}(\boldsymbol{E}^x_{i,[1:j+1]})$. In this way, the diagnosed diseases at the early period are explicitly encoded into the i-th patient's current physician state $\boldsymbol{E}^v_{i,j}$. Then, we further apply an attention mechanism across $\boldsymbol{E}^s_i = [\boldsymbol{E}^v_{i,1}, \boldsymbol{E}^v_{i,2}, \ldots, \boldsymbol{E}^v_{i,N_i-1}]$, and obtain the representation of each sequence as $\boldsymbol{s}_i = \text{softmax}(\boldsymbol{E}^s_i \boldsymbol{W}_s)\boldsymbol{E}^s_i$, where $\boldsymbol{W}_s$ is learnable parameters.

3.3.2 Inference with CNF. With the sequence representation $\boldsymbol{s}_i$ learned, it is widely adopted to leverage the reparameterization trick [12, 15] to express the random variable, i.e., $\boldsymbol{z}_i \sim \mathcal{N}(\mu_i, \text{diag}(\sigma_i^2))$ as follows:

$$\begin{aligned} [\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i] &= \text{Encoder}(\boldsymbol{s}_i; h_\theta), \\ \boldsymbol{z}_i &= \boldsymbol{\mu}_i + \epsilon \odot \boldsymbol{\sigma}_i, \quad \epsilon \sim \mathcal{N}(0, \boldsymbol{I}). \end{aligned} \tag{4}$$

Although it is possible to use a simple Gaussian prior over the real posterior distribution $q_\phi(z_i|\boldsymbol{\tau}_i)$, 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_ϕ. 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 $\boldsymbol{z}_i$ and its log densities to transform $q(z_i|\boldsymbol{\tau}_i)$ in a continuous way. Given a variable $\boldsymbol{z}_i(0)$ with known probability distribution $p(\boldsymbol{z}_i(0))$ (e.g., Gaussian in Eq. 4) and differential function β_ζ that is uniformly Lipschitz continuous in both $\boldsymbol{z}_i$ and step ψ, we have:

$$\frac{d\boldsymbol{z}_i(\psi)}{d\psi} = \beta_\zeta(\boldsymbol{z}_i(\psi), \psi), \tag{5}$$

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

$$\frac{d \log q(\boldsymbol{z}_i(\Psi)|\boldsymbol{\tau}_i)}{d\psi} = -\text{Tr}\left(\frac{\partial \beta_\zeta(\boldsymbol{z}_i(\psi), \psi)}{\partial \boldsymbol{z}_i(\psi)}\right), \tag{6}$$

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

Then, the latent variables $\boldsymbol{z}_i(\Psi)$ after step Ψ and its log densities $\log q(\boldsymbol{z}_i(\Psi)|\boldsymbol{\tau})$ can be computed as:

$$\begin{aligned} \mathrm{z}_i(\Psi) &= \mathrm{z}_i(0) + \int_0^\Psi \beta_\zeta\left(\mathrm{z}_i(\psi), \psi\right) d\psi, \\ \log q_\phi\left(\mathrm{z}_i(\Psi)|\boldsymbol{\tau}_i\right) &= \log q_\phi\left(\mathrm{z}_i(0)|\boldsymbol{\tau}_i\right) - \int_0^\Psi \text{Tr}\left(\frac{\partial \beta_\zeta}{\partial \mathrm{z}_i(\psi)}\right) d\psi, \end{aligned} \tag{7}$$

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

With $\log q_\phi\left(\mathrm{z}_i(\Psi)|\boldsymbol{\tau}_i\right)$ in Eq. 7, the *matching* part in Eq. 2 can be calculated based on CNF as follows:

$$\begin{aligned} \mathcal{L}_m = \mathbb{E}_{q_\phi(\boldsymbol{z}_i|\boldsymbol{\tau}_i)} \log \frac{q_\phi\left(\boldsymbol{z}_i(0)|\boldsymbol{\tau}_i\right)}{p\left(\boldsymbol{z}_i(0)\right)} = \mathbb{E}_{q_\phi(\boldsymbol{z}_i|\boldsymbol{\tau}_i)} \log p\left(\boldsymbol{z}_i(\Psi)\right) \\ - q_\phi\left(\boldsymbol{z}_i(0)|\boldsymbol{\tau}_i\right) + \int_0^\Psi \text{Tr}\left(\frac{\partial \beta_\zeta(\boldsymbol{z}_i(\psi), \psi)}{\partial \boldsymbol{z}_i(\psi)}\right) d\psi]. \end{aligned} \tag{8}$$

3.4 Decoder

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

$q_\phi(\boldsymbol{z}_i(0)|\mathcal{S}_i)$. To this end, the goal of decoder g_ν in Eq. 3 can be simplified as follows:

$$\begin{aligned}&\log p(\boldsymbol{\tau}_{i,j+1}|\boldsymbol{\tau}_{i,[1:j]}, \mathbf{z}_i(\Psi))\\ &= -\left[\log p\left(\hat{\boldsymbol{d}}_{i,j+1}|\boldsymbol{\tau}_{i,[1:j]}, \mathbf{z}_i(\Psi)\right) + \log p\left(\hat{T}_{i,j+1}|\boldsymbol{\tau}_{i,[1:j]}, \mathbf{z}_i(\Psi)\right)\right]\\ &= \text{Decoder}(\boldsymbol{\tau}_{i,[1:j]}, \mathbf{z}_i(\Psi); g_\nu),\end{aligned} \tag{9}$$

where $\boldsymbol{\tau}_{i,j+1} = \left(\boldsymbol{d}_{i,j+1}, T_{i,j+1}\right)$ includes the diagnosis $\boldsymbol{d}_{i,j+1}$ and the exact diagnosis time $T_{i,j+1}$. In particular, we first obtain the temporal context representation $\boldsymbol{\Gamma}_{i,j} \in \mathbb{R}^{D\times 1}$ as follows:

$$\boldsymbol{\Gamma}_{i,j} = \text{MLP}([\boldsymbol{\tau}_{i,[1:j]}, \boldsymbol{z}_i(\Psi)]), \tag{10}$$

where $\boldsymbol{\tau}_{i,[1:j]}$ is the input and $\boldsymbol{z}_i(\Psi)$ is the latent variable in Eq. 7.

Diagnosis prediction layer aims to predict the diagnosed diseases of each patient. Given the latent variable $\boldsymbol{z}_i(\Psi)$ in Eq. 7, the probability of the predictive disease $\hat{\boldsymbol{d}}_{i,j+1}$ is calculated as follows:

$$p\left(\hat{\boldsymbol{d}}_{i,j+1}|\boldsymbol{\Gamma}_{i,j}\right) = \text{softmax}(\boldsymbol{\Gamma}_{i,j}). \tag{11}$$

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

$$\lambda_i(t) = \exp\left(\boldsymbol{\Gamma}_{i,j} \cdot \boldsymbol{v}^t + \alpha\left(t - T_{i,j}\right) + \lambda_0\right), \tag{12}$$

where $\boldsymbol{v}^t \in \mathbb{R}^{D\times 1}$ and λ_0 are scalars that denote the basic intensity of the next diagnosis. The first term $\boldsymbol{\Gamma}_{i,j} \cdot \boldsymbol{v}^t$ calculates the accumulative effect among the past diseases. The second term $\alpha(t - T_{i,j})$ denotes the evolutional 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:

$$p\left(\hat{T}_{i,j+1} = T_{i,j+1}|\boldsymbol{\Gamma}_{i,j}\right) = f_i(T_{i,j+1} - T_{i,j}) = f_i(\Delta T_{i,j+1}), \tag{13}$$

where $f_i(t) = \lambda_i(t)\exp\left(-\int_{T_{i,j}}^{t}\lambda(\varepsilon)d\varepsilon\right)$. 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:

$$\begin{aligned}\min \mathcal{L}_{pred} = -\sum_{i=1}^{|\mathcal{T}^{tr}|}\sum_{j=1}^{N_{Q_i}-1}&\left[\log p(\hat{\boldsymbol{d}}_{i,j+1}|\boldsymbol{\Gamma}_{i,j})\right.\\ &\left.+\log p(\hat{T}_{i,j+1} = T_{i,j+1}|\boldsymbol{\Gamma}_{i,j}\right] - \mathcal{L}_m,\end{aligned} \tag{14}$$

where $p(\hat{\boldsymbol{d}}_{i,j+1}|\boldsymbol{\Gamma}_{i,j})$ is calculated in Eq. 11, $p(\hat{T}_{i,j+1} = T_{i,j+1}|\boldsymbol{\Gamma}_{i,j})$ is calculated in Eq. 13, and the matching loss $\mathcal{L}_m$ is calculated in Eq. 8. $\boldsymbol{d}_{j+1}$ and $T_{i,j+1}$ 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 hierarchically 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:

$$\boldsymbol{d}_j = \boldsymbol{E}_j^1 \oplus \boldsymbol{E}_j^2 \oplus \cdots \oplus \boldsymbol{E}_j^l \oplus \cdots \oplus \boldsymbol{E}_j^L \in \mathbb{R}^{D\times 1}, \tag{15}$$

where $\oplus$ denotes the concatenation operation and $\boldsymbol{E}_j^l$ 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 Υ), 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 patient-disease matrix Υ, we can obtain the concurrent frequency of two diseases. Specifically, we first multiply $\Upsilon^T\Upsilon$ 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 diagnoised 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 $\mathcal{G}$, it is straightforward to reorganize the embedding space by regularizing the diseases insides 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 patents (i.e., 272 and 401) to be close to each other.

In light of this, we propose to disentangle the disease embeddings to $\boldsymbol{E}_j^l$ 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:

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

where $\|\cdot\|^2$ measure the Euclidean distance and $\boldsymbol{E}_i^l$ is the i-th disease's embedding at level l. l_i denotes d_i's acutal 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 $\boldsymbol{u}_i$, the patient embeddings can inherit both the hierarchical and syndromic relations among diseases. Since $\boldsymbol{u}_i$ 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 $\boldsymbol{u}_i$ is given as follows:

$$\boldsymbol{u}_i = \frac{\sum_i v_i \boldsymbol{d}_j}{\sum_i v_i} = \boldsymbol{E}_i^{u,1} \oplus \boldsymbol{E}_i^{u,1} \cdots \oplus \boldsymbol{E}_i^{u,L}, \tag{17}$$

where $v_i \in \Upsilon$ is from patient-disease matrix and $\boldsymbol{E}_i^{u,l}$ represents the l-th level feature of patient u_i.

Then, since u_i inherits the hierarchical property of disease by aggregation, we propose to perform hierarchical subytping, where the patient subtypes are dentoed 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 $\boldsymbol{H} = [\boldsymbol{H}_1^1, \boldsymbol{H}_2^1, \ldots \boldsymbol{H}_C^1, \ldots, \boldsymbol{H}_C^L]$. To measure the importance of different subtypes in $\mathcal{H}$ for u_i, we calculate the attention score $\alpha_{i,j}^l$ between patient u_i and the subtype $\boldsymbol{H}_j^l$ at the l-th level as:

$$\alpha_{i,j}^l = \|\boldsymbol{E}_i^{u,l} - \boldsymbol{H}_j^l\|^2 / \sum_{j'} \|\boldsymbol{E}_i^{u,l} - \boldsymbol{H}_{j'}^l\|^2. \tag{18}$$

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

$$\boldsymbol{P}_i = \text{MLP}(\boldsymbol{u}_i + \sum_{l=1}^{L} \sum_{j=1}^{lC} \alpha_{i,j}^l \boldsymbol{H}_j^l). \tag{19}$$

Algorithm 1: Training process of `MetaCare++`

Data: Training task set $\mathcal{T}^{tr}$; the disease set $\mathcal{D}$, Hyperparameters: K, C, and ω.
Result: Parameters in embedding layer; h_θ; $\boldsymbol{H}$; g_{v_i}.

```
Initialize all model parameters;
while not converged do
    for all τ_i ∈ T^tr do
        Construct support sequence S_i and query sequence Q_i from τ_i as Section 3.1;
        Generate q_φ(z_i(Ψ)|τ_i) via encoder h_θ in Eq.7;
        Predictions on Q_i via personalized decoder g_{v_i} in Eq. 21;
        Generate q_φ(z_i(Ψ)|S_i) via encoder h_θ in Eq.7 ;
        Calculate prediction loss L_pred in Eq. 14 ;
    Calculate total loss in Eq. 22, including L_pred in Eq. 14, L_d−subtyping in Eq. 16, and L_p−subtyping in Eq. 20;
    Update patient embedding u, disease embedding d, h_θ, and g_{v_i};
```

Table 1: Statistics of the datasets used in our experiments.

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

Similar to the regularization 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 $\boldsymbol{H}_j^l$ at different levels l as follows:

$$\mathcal{L}_{p-subtpying} = \sum_{l=1}^{L} \sum_{j=1}^{lC} \|\boldsymbol{H}_j^l - \frac{1}{n_j^l} \sum_{u_i \in \mathcal{H}_j^l} \boldsymbol{E}_i^{u,l}\|^2, \tag{20}$$

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

Finally, we learn a personalized decoder g_{v_i} to adapt only the personalized parameters from general decoder g_v as follows:

$$g_{v_i} = \boldsymbol{P}_i \circ g_v, \tag{21}$$

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:

$$\min \mathcal{L}_{pred} + \omega(\mathcal{L}_{d-subtyping} + \mathcal{L}_{p-subtyping}), \tag{22}$$

where $\mathcal{L}_{pred}$ is calculated in in Eq. 14 for both diagnosis and diagnosis times prediction tasks. ω is a hyperparameter to control the weight for the subtyping of diseases (i.e., $\mathcal{L}_{d-subtyping}$ in Eq. 16) and patients($\mathcal{L}_{p-subtyping}$ 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. For diagnosis prediction, we use Recall@K and NDCG@K metrics. Intuitively, the Recall metric considers whether the ground-truth is ranked amongst the top K diagnoses while the NDCG metric is a position-aware ranking metric. For diagnosis time prediction, we use the MAE metric, which measures the mean absolute error between the predicted time and the ground truth.

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] propose 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/

Table 2: Experimental results on two benchmark datasets. The best performance is in boldface and the second runners are underlined. `MetaCare++` achieves the best performance on all datasets, where * denotes a significant improvement according to the Wilcoxon signed-rank test.

Method	Recall@5	NDCG@5	Recall@10	NDCG@10	MAE	Recall@5	NDCG@5	Recall@10	NDCG@10	MAE
	MIMIC-III					eICU				
RATAIN	0.0919	0.2899	0.1388	0.2663	0.1626	0.3732	0.3942	0.5054	0.4106	0.0309
Dipole	0.0724	0.2579	0.1257	0.2587	0.1943	0.3429	0.3618	0.4733	0.3921	0.0335
Timeline	0.0853	0.2636	0.1282	0.2622	0.1710	0.3644	0.3785	0.4936	0.4035	0.0316
GRAM	0.1045	0.3265	0.1702	0.2845	0.1577	0.3952	0.4066	0.5165	0.4327	0.0284
KAME	0.1033	0.3218	0.1696	0.2793	0.1609	0.3878	0.4012	0.5097	0.4218	0.0282
MHM	0.1056	0.3289	0.1756	0.3054	0.1504	0.4144	0.4192	0.5322	0.4416	0.0279
TAdaNet	0.1094	0.3316	0.1792	0.3130	0.1493	0.4161	0.4227	0.5352	0.4431	0.0268
CGL	0.1174	0.3438	0.1793	0.3160	0.1319	0.4159	0.4291	0.5368	0.4532	0.0254
MeLU	0.0945	0.3208	0.1507	0.2887	0.1443	0.3996	0.4118	0.5184	0.4473	0.0268
MAMO	0.1004	0.3354	0.1548	0.3147	0.1408	0.4143	0.4219	0.5236	0.4452	0.0270
TaNP	0.1012	0.3395	0.1587	0.3151	0.1322	0.4170	0.4324	0.5339	0.4436	0.0257
`MetaCare`	0.1195	0.3478	0.1831	0.3214	0.1293	0.4245	0.4372	0.5437	0.4582	0.0251
`MetaCare++`	**0.1296***	**0.3725***	**0.1920***	**0.3486***	**0.1107***	**0.4468***	**0.4551***	**0.5640***	**0.4897***	**0.0224***

(RETAIN[2], GRAM [3], MHM [4], CGL [5], MeLU [6], MAMO [7], and TaNP [8]). We follow the original settings suggested by the authors to train all baseline models. For fair comparisons, we make sure that only patient-disease interactions and the ICD9 information are included.

For models under the meta-learning setting, we follow the same setting as [18] and split the dataset into training and testing with a ratio of 1:1. Specifically, we split all patients $\mathcal{U}$ into two disjoint sets: the visits with training patient set (i.e., $\mathcal{T}^{Tr}$) and the visits with test (cold-start) patient set (i.e., $\mathcal{T}^{Te}$). Among $\mathcal{T}^{Tr}$ and $\mathcal{T}^{Te}$, we split the first 30% visits to serve as the support sequenece and the last 70% as the query sequenece according to diagnosis time. For the other models, we train them with the same data (all visits in $\mathcal{T}^{Tr}$ and support sets in $\mathcal{T}^{Te}$) for fair comparisons. We tune all hyperparameters through grid search. In particular, learning rate in {1e-5, 5e-5, 1e-4, 5e-4, 1e-3}, the number for splitting diagnosis sets K in {5, 10, 15, 20, 25}, and weight ω in {0, 0.001, 0.01, 0.1, 1.0}. We set the embedding dimension D to 64 for all compared methods on both MIMIC-III and eICU. The batch size is set to 32. We also carefully tune the hyperparameters of baselines on the validation set as suggested in the original papers to achieve their best performance.

5.2 Overall Performance Comparison (RQ1)

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 score on MIMIC-III and eICU datasets with K={5,10}. We have the following observations.

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.

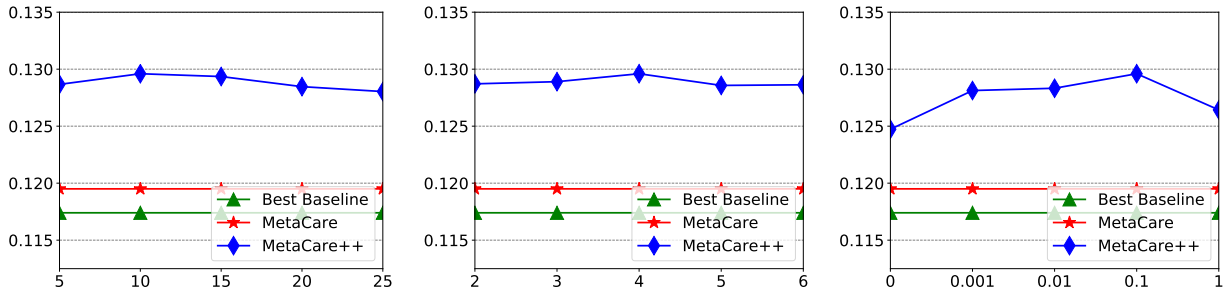
In particular, by considering latent temporal relations among patient visits and collaboratively learning representations of patients and diseases, CGL performs better than TaNP in many cases. Compared with CGL, `MetaCare++` not only takes the data sparsity problem into consideration but also explicitly 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 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

[2]https://github.com/mp2893/retain
[3]https://github.com/mp2893/gram
[4]https://github.com/qxiaobu/MHM
[5]https://github.com/LuChang-CS/CGL
[6]https://github.com/hoyeoplee/MeLU
[7]https://github.com/dongmanqing/Code-for-MAMO
[8]https://github.com/IIEdm/TaNP.



(a) Varying K (b) Varying C (c) Varying ω

Figure 3: Performance regarding Recall@5 of the best baseline, MetaCare, and MetaCare++ on the MIMIC-III dataset.

Table 3: Ablation analysis of our proposed MetaCare++ on the MIMIC-III and eICU datasets.

Method	Recall@5	NDCG@5	MAE
	MIMIC-III		
Meta	0.0942	0.3225	-
Meta + CNF	0.1064	0.3328	-
MetaCare	0.1195	0.3478	0.1293
MetaCare + Diag-Disease	0.1254	0.3602	0.1175
MetaCare++	**0.1296**	**0.3725**	**0.1107**
	eICU		
Meta	0.4099	0.4187	-
Meta + CNF	0.4165	0.4266	-
MetaCare	0.4245	0.4372	0.0251
MetaCare + Diag-Disease	0.4407	0.4498	0.0228
MetaCare++	**0.4468**	**0.4551**	**0.0224**

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.

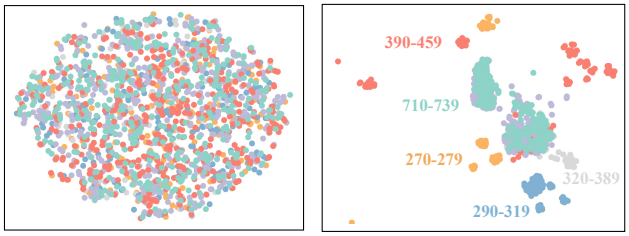
Furthermore, the performance gains of MetaCare + Diag-Disease over MetaCare ranges from 2.88% (achieved in NDCG@5 on eICU) to 9.90% (achieved in MAE on MIMIC-III). The result shows that: (1) the explicitly modeling hierarchical and syndromic relations among diseases can further improve the performance 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 ω.

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) ω 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 ω will cause the tag embeddings likely to be spread out, while too large ω will likely cause the model to overfit. The optimal λ value on MIMIC-III is about 0.1. Note that, MetaCare++ is reasonably sensitive to ω. In the range of [0.1, 1], the optimal ω can be obtained by slight tuning.



(a) MetaCare (b) MetaCare++

Figure 4: Disease embeddings learned by MetaCare and MetaCare++ on the MIMIC-III dataset, where the different color of numbers are the code of different types of diseases.

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 exampled 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 lipoid metabolism, the conoccurent 272.0 and 401.0 leads to a risk of *Metabolic syndrome* rather than pure *Cardiovascular Syndrom* 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.

	Visit	Subtype	Prediction
Lisa	431;401.0;. . .	*Cardiovascular Syndrome*	414.01;. . .
Jack	435.9; 401.0;. . .	*Cardiovascular Syndrome*	427.11;. . .
Linda	272.0; 401.0;. . .	*Metabolic syndrome*	790.5;. . .

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

[1] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. De Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, 2016.
[2] T. Bai, S. Zhang, B. L. Egleston, and S. Vucetic. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 43–51, 2018.
[3] R. T. Chen and D. Duvenaud. Neural networks with cheap differential operators. *arXiv preprint arXiv:1912.03579*, 2019.
[4] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366*, 2018.
[5] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.
[6] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, 2016.
[7] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 787–795, 2017.
[8] M. Dong, F. Yuan, L. Yao, X. Xu, and L. Zhu. Mamo: Memory-augmented meta-optimization for cold-start recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 688–697, 2020.
[9] Z. Du, X. Wang, H. Yang, J. Zhou, and J. Tang. Sequential scenario-specific meta learner for online recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2895–2904, 2019.

[10] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.

[11] W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In *ICLR*, 2019.

[12] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[13] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 2016.

[14] H.-C. Kao, K.-F. Tang, and E. Chang. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[15] D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. *Advances in Neural Information Processing Systems*, 2015.

[16] H. Lee, J. Im, S. Jang, H. Cho, and S. Chung. Melu: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1073–1082, 2019.

[17] A. Likas, N. Vlassis, and J. J. Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[18] X. Lin, J. Wu, C. Zhou, S. Pan, Y. Cao, and B. Wang. Task-adaptive neural process for user cold-start recommendation. In *The World Wide Web Conference*, 2021.

[19] L. Liu, Z. Liu, H. Wu, Z. Wang, J. Shen, Y. Song, and M. Zhang. Multi-task learning via adaptation to similar tasks for mortality prediction of diverse rare diseases. In *AMIA Annual Symposium Proceedings*.

[20] C. Lu, C. K. Reddy, P. Chakraborty, S. Kleinberg, and Y. Ning. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. In *IJCAI*, 2021.

[21] Q. Lu, T. H. Nguyen, and D. Dou. Predicting patient readmission risk from medical text via knowledge graph enhanced multiview graph convolution. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1990–1994, 2021.

[22] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1903–1911, 2017.

[23] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *CIKM*, 2018.

[24] L. Ma, C. Zhang, Y. Wang, W. Ruan, J. Wang, W. Tang, X. Ma, X. Gao, and J. Gao. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[25] E. Mathieu and M. Nickel. Riemannian continuous normalizing flows. *arXiv preprint arXiv:2006.10605*, 2020.

[26] A. Obamuyide and A. Vlachos. Model-agnostic meta-learning for relation classification with limited supervision. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5873–5879, 2019.

[27] X. Peng, G. Long, T. Shen, S. Wang, Z. Niu, and C. Zhang. Mimo: Mutual integration of patient journey and medical ontology for healthcare representation learning. *arXiv preprint arXiv:2107.09288*, 2021.

[28] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 2018.

[29] Z. Qian, A. Alaa, A. Bellot, M. Schaar, and J. Rashbass. Learning dynamic and personalized comorbidity networks from event data using deep diffusion processes. In *AISTATS*, 2020.

[30] Z. Qian, W. R. Zame, M. van der Schaar, L. M. Fleuren, and P. Elbers. Integrating expert odes into neural odes: Pharmacology and disease progression. In *Advances in Neural Information Processing Systems*, 2021.

[31] Z. Qiao, Z. Zhang, X. Wu, S. Ge, and W. Fan. Mhm: Multi-modal clinical data based hierarchical multi-label diagnosis prediction. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1841–1844, 2020.

[32] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali. Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data. *Medical care*, pages 1130–1139, 2005.

[33] E. Rocheteau, P. Liò, and S. Hyland. Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 58–68, 2021.

[34] L. Song, C. W. Cheong, K. Yin, W. K. Cheung, B. C. Fung, and J. Poon. Medical concept embedding with multiple ontological representations. In *IJCAI*, 2019.

[35] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[36] Q. Suo, J. Chou, W. Zhong, and A. Zhang. Tadanet: Task-adaptive network for graph-enriched meta-learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1789–1799, 2020.

[37] X. Teng, S. Pei, and Y.-R. Lin. Stocast: Stochastic disease forecasting with progression uncertainty. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[38] J. Whang, E. Lindgren, and A. Dimakis. Composing normalizing flows for inverse problems. In *International Conference on Machine Learning*, 2021.

[39] C. Xu, Y. Fu, C. Liu, C. Wang, J. Li, F. Huang, L. Zhang, and X. Xue. Learning dynamic alignment via meta-filter for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[40] G. Yang, X. Huang, Z. Hao, M.-Y. Liu, S. Belongie, and B. Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[41] M. Ye, S. Cui, Y. Wang, J. Luo, C. Xiao, and F. Ma. Medpath: Augmenting health risk prediction via medical knowledge paths. In *The World Wide Web Conference*, 2021.

[42] C. Yu, J. Han, H. Zhang, and W. Ng. Hypernymy detection for low-resource languages via meta learning. In *ACL*, 2020.

[43] X. S. Zhang, F. Tang, H. H. Dodge, J. Zhou, and F. Wang. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2487–2495, 2019.

[44] K. Zheng, W. Wang, J. Gao, K. Y. Ngiam, B. C. Ooi, and W. L. J. Yip. Capturing feature-level irregularity in disease progression modeling. In *CIKM*, 2017.

[45] F. Zhou, L. Li, K. Zhang, G. Trajcevski, F. Yao, Y. Huang, T. Zhong, J. Wang, and Q. Liu. Forecasting the evolution of hydropower generation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2861–2870, 2020.

[46] F. Zhou, Z. Wen, K. Zhang, G. Trajcevski, and T. Zhong. Variational session-based recommendation using normalizing flows. In *The World Wide Web Conference*, 2019.