

Original Research

Knowledge-aware multi-center clinical dataset adaptation: Problem, method, and application

Jiebin Chu^a, Jinbiao Chen^a, Xiaofang Chen^a, Wei Dong^b, Jinlong Shi^c, Zhengxing Huang^{a,*}^a College of Biomedical Engineering and Instrument Science, Zhejiang University, China^b Department of Cardiology, Chinese PLA General Hospital, China^c Department of Medical Innovation Research, Medical Big Data Center, Chinese PLA General Hospital, China

ARTICLE INFO

Keywords:

Clinical dataset adaptation
 Knowledge graph
 Clinical outcome prediction
 Heart failure
 Representation learning

ABSTRACT

Adaptable utilization of clinical data collected from multiple centers, prompted by the need to overcome the shifts between the dataset distributions, and exploit these different datasets for potential clinical applications, has received significant attention in recent years. In this study, we propose a novel approach to this task by infusing an external knowledge graph (KG) into multi-center clinical data mining. Specifically, we propose an adversarial learning model to capture shared patient feature representations from multi-center heterogeneous clinical datasets, and employ an external KG to enrich the semantics of the patient sample by providing both clinical center-specific and center-general knowledge features, which are trained with a graph convolutional autoencoder. We evaluate the proposed model on a real clinical dataset extracted from the general cardiology wards of a Chinese hospital and a well-known public clinical dataset (MIMIC III, pertaining to ICU clinical settings) for the task of predicting acute kidney injury in patients with heart failure. The achieved experimental results demonstrate the efficacy of our proposed model.

1. Introduction

Several large clinical datasets, particularly those collected from different clinical research centers, include a large number of participants from diverse geographic locations and center-specific features [1,2]. Utilizing clinical data collected from multiple centers to prove or disprove a hypothesis across various clinical settings is essential to improve patient therapy and care, provide high quality healthcare management, and conduct effective clinical research [3]. In several cases, multi-centers clinical datasets have different genetic, environmental, and ethnical distribution, and the efficacy of learning from these data will vary significantly among the centers [4,5]. As indicated in several large-scale studies [6,7], machine learning models trained on data collected from one clinical center cannot be reliably deployed in the other clinical settings owing to the distribution differences between the training and evaluation datasets, known as the dataset shift in the machine learning field [8]. Clinical dataset adaptation, as a special case of data fusion, is a process of integrating multiple data sources with similar types of real clinical objects in a consistent, accurate, and useful way [6,9]. With the potential to solve the dataset shift problem observed while utilizing multi-center datasets, clinical dataset adaptation is

gaining increasing attention in the medical informatics field. In this paper, we propose a new clinical dataset adaptation methodology to overcome dataset shift and allow the application of the model trained on a labeled source dataset to a unlabeled target dataset.

Research in this field has proposed diverse approaches to explore the huge potential of heterogeneous multi-center data for essential clinical applications, such as treatment effect estimation [10–12] and the clinical outcome prediction task addressed in this study [13–15]. Notably, clinical outcomes, e.g., length of stay (LOS), readmission time, and discharge type, have been recognized as critical and essential indicators of medical service delivery in clinical practice. As essential implications in health service delivery, clinical outcome prediction plays an important role for both patients and healthcare providers [16]. Notably, substantial efforts have been undertaken to transform heterogeneous clinical data into one common format (data model [17,18]) using standard terminologies/lexicons, and then perform systematic analyses by employing standard analytic routines as well as pre-trained data mining and machine learning models based on the common format [19,20]. Although valuable, these approaches are limited owing to the shift in data distribution among multiple centers [21,22].

As a remedy, traditional studies, especially those focused on the

* Corresponding author.

E-mail address: zhengxinghuang@zju.edu.cn (Z. Huang).<https://doi.org/10.1016/j.jbi.2021.103710>

Received 23 September 2020; Received in revised form 5 February 2021; Accepted 6 February 2021

Available online 11 February 2021

1532-0464/© 2021 Elsevier Inc. This article is made available under the Elsevier license (<http://www.elsevier.com/open-access/userlicense/1.0/>).

treatment effect estimation tasks, have adopted propensity score matching (PSM) to map the source and target clinical datasets. Thus, knowledge learned from one dataset can be applied to the other datasets [23–25], especially when it incorporates the learned mapping between datasets extracted from multiple clinical centers.

Recently, with the continued success of deep learning, representation learning has emerged as an alternative solution to address the biased data distribution problem [26]. Specifically, a few efforts have been made to adopt deep learning tactics, such as deep neural multi-view network [27] and adversarial learning [28–30], to extract latent shared patient features from heterogeneous clinical data.

Despite the unique advantages provided by both propensity score analysis and representation learning, they share the same critical limitation, that is, the lack of ability to retain center-specific patient features properly during learning. Defined as the difference in their joint distribution of covariates and labels, dataset shift indicates that some patient features have variant influence on outcomes in different clinical settings. This part of the features is called center-specific patient features. For example, angiogram is an important imageological examination for heart failure (HF) patients to check their blood perfusion state. The contrast agent used in this examination is excreted by the kidney. For patients in general wards, angiogram rarely causes any adverse impact on the kidney. However, for patients in intensive care units (ICUs), who are in a much worse condition, using the contrast agent brings extra burden to the kidney and increases the risk of acute kidney injury (AKI). Existing methods can hardly capture the variational relation between the center-specific features and outcomes in datasets collected from different clinical settings. As a result, the performance of clinical applications based on multi-center clinical data is inevitably degraded.

To alleviate this problem, we propose a Knowledge-Aware Multi-center clinical dataset Adaptation model (KAMA) and validate it in a common application, i.e., clinical outcome prediction. In detail, we considered learning from the dataset of each clinical center as a single task and adopted the adversarial learning strategy to extract the common underlying and center-invariant patient features, and then used them for further clinical applications. Meanwhile, we attempted to improve the learning model by infusing an external knowledge graph (KG) into the model. Specifically, the proposed KAMA learns the KG embeddings using a graph convolution network (GCN). The clinical center-specific KG embeddings can then be extracted and fed to a deep neural network model, which comprises (1) a center discriminator to distinguish patient samples between the source and target clinical centers so that center-invariant features can be adversarially captured, and (2) an application module (concretely, a clinical outcome predictor in our case study) that is used during training and testing. Notably, the parameters of the underlying deep neural network are optimized to minimize the loss of the outcome prediction and maximize that of the center discriminator. Thus, the proposed model can perform clinical outcome prediction reliably in diverse clinical settings across multiple clinical centers.

We validated the proposed model in a typical clinical task, i.e., clinical outcome prediction. Specifically, we collected a real clinical dataset from the general wards of cardiology department at the Chinese PLA general hospital, and a widely used public clinical dataset MIMIC III related to the ICU setting. We extracted the samples of patients with HF and considered the occurrence of AKI as their clinical outcome. To evaluate the effectiveness of clinical dataset adaptation, we assume that the labels of the target dataset are absent in training process. In this scenario, we trained the models on the labeled source dataset while the features of unlabeled target dataset are also utilized, and then evaluated the learned models on the target dataset. The better performance obtained in this scenario indicates the efficacy on clinical dataset adaptation.

We compared our model with state-of-the-art models, and the experimental results demonstrated that our model significantly outperformed them on the experimental dataset, thus validating our claim

that external knowledge can be well exploited to aid the task of multi-center clinical dataset adaptation, and facilitate downstream tasks, such as clinical outcome prediction.

In Section 2, we present a brief review of related work. In Section 3, we define the problem and describe preliminary knowledge used in this study. In Section 4, we present our proposed model in detail. Section 5 describes the experimental setup and reports the results to verify the superiority of our model. Finally, Section 6 concludes this paper.

2. Related work

As a special case of data fusion, clinical dataset adaptation is a vital process to overcome the shift in datasets collected from multiple centers. Conventional data fusion aims to merge such datasets by mapping them on a unified data model/schema [17,18]. In the era of big data, there are multiple datasets generated in different clinical settings, which are implicitly related to the same clinical task (e.g., diagnosis, outcome prediction). Literally, patient samples from multi-center clinical datasets, even with the same first diagnosis or disease type, have different genetic, environmental, and ethnical distribution [6]. Thus, we cannot merge them in a straightforward manner, with schema mapping. Instead, we are required to exploit the method of obtaining robust representations of patients from multi-center clinical datasets to facilitate the downstream clinical applications [31,32]. From this point of view, this is more about knowledge fusion instead of mapping via a standard data model, and thus significantly differentiates between traditional data fusion and multi-center clinical dataset adaptation.

Studies on this topic can be broadly categorized into two methodologies: (1) instance matching and (2) representation learning. The most widely used instance-matching method is PSM, which builds a logistic regression to estimate the probabilities of patients to be assigned with the treatment. The probabilities are called propensity scores. Patients with close scores are considered similar to one another. PSM searches the patients in the control group to match the patients in the treated group in accordance with the propensity scores. For example, Valgimigli et al. [33] used propensity score adjustment and matching to investigate the effect of transfemoral versus transradial intervention in acute myocardial infarction on a multi-center dataset. By using instance-matching methods such as PSM, a matched set comprising patient samples collected from the source and target clinical centers and bearing a similar feature distribution can be generated to approximate a randomized controlled trial (RCT), eliminating several problems seen with observational data analysis [23–25,34]. Despite its popularity, most instance-matching approaches adopt linear shallow models, which makes it difficult to learn the shared feature representations from nonlinear heterogeneous clinical data, usually resulting in a complex model [35,36]. More importantly, instance-matching approaches lack the ability to model clinical setting-specific information. Recent studies have attempted to focus on developing deep representation learning techniques for multi-center clinical data adaptation [31,32,37]. Different from matching patient samples at the instance level, in deep representation learning, the (dis)similarity between distributions is measured, which is accomplished by updating the parameters of neural networks to generate modified patient feature representation instead of operating the original input with geometric transformation [38,39]. Notably, as a popular approach in this big data era, deep representation learning can provide a convenient way to combat the data distribution gap between multi-center clinical datasets, which is essential for tasks in healthcare intelligence such as diagnosis prediction, cohort selection, and treatment effect estimation. For example, Zellinger et al. [40] proposed a representation learning method with a new distance function called central moment discrepancy (CMD) for regularization to minimize the discrepancy between domain-specific latent feature representations. Pokharel et al. proposed a novel representation for information in structured health records that aims to explicitly model temporality, variety, heterogeneity, irregularity, and sparsity [41]. In [42], Ma et al.

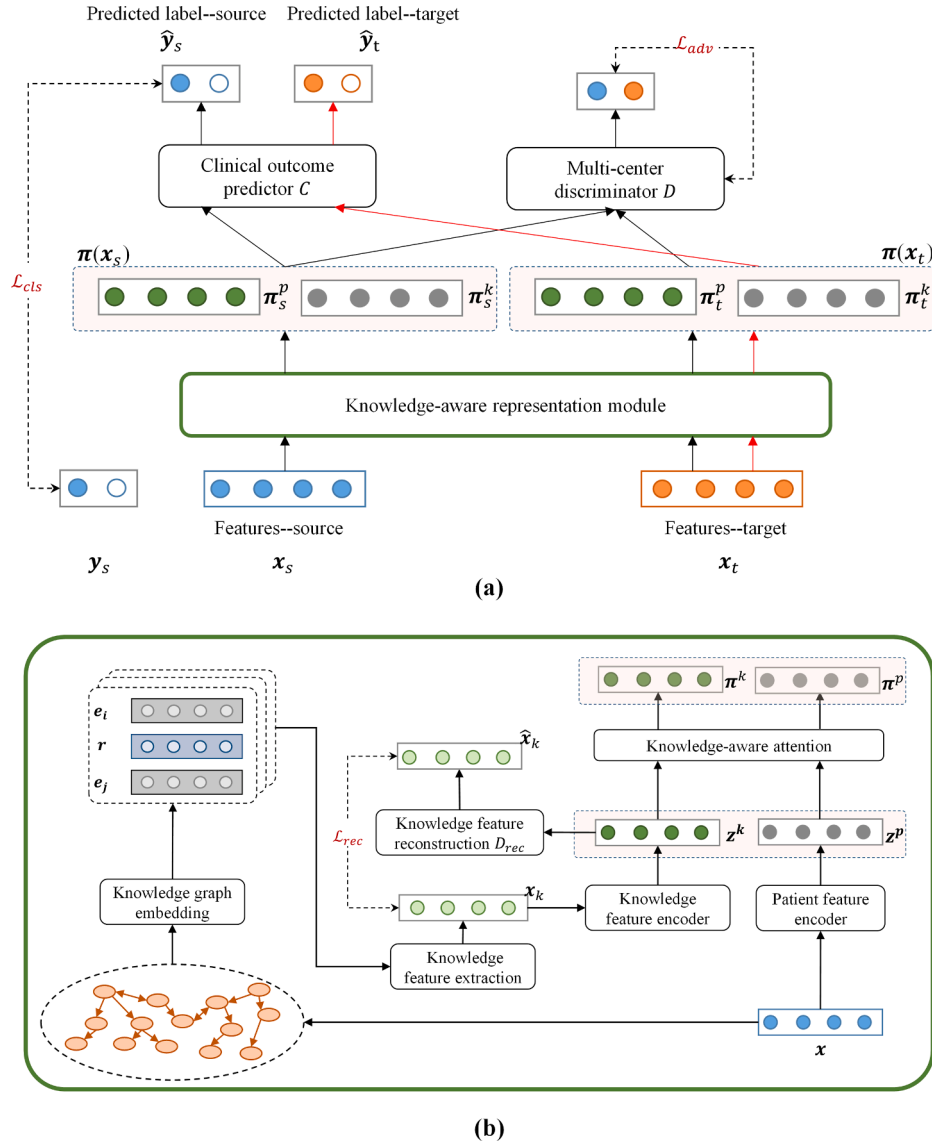


Fig. 1. The proposed knowledge-aware multi-center clinical dataset adaptation model for outcome prediction. (a) Schema of the proposed model. The flow lines marked in black show the training process and the flow lines marked in red show the testing process on the target dataset; (b) Knowledge-aware representation module. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

presented a new phenotype network representation method learned from PubMed. Specifically, their model takes semantic representations of phenotypes as input and predicts synonymous relationships between phenotype concepts from different terminology databases. In [12], Chen et al. formulated treatment effect estimation as a representation learning problem and proposed learning the latent representations through multi-task deep learning of electronic health records. Based on the learned representations, the counterfactual treatment effect can be estimated in an interpretable manner [10–12]. Recently, generative adversarial network (GAN) has achieved popularity in representation learning with its unique idea of mapping distributions with adversarial training strategy. GAN builds two neural networks, called generator and discriminator, respectively, to contest with each other. During the contesting process, the generator learns to map the distribution of noise signal input to be to that of true samples [28]. Following this idea, recent studies adopted adversarial learning strategy to map the samples from different data sources to latent representations with similar distributions so that the shared information is encoded into the representations. For example, Huang et al. proposed a multi-task adversarial neural network model to predict major adverse cardiac events for acute coronary

syndrome (ACS) patients with the three different subtypes, based on a large volume of heterogeneous electronic health records [29]. Specifically, they incorporated adversarial learning into the model to prevent both the shared and private latent features of each ACS subtype from interfering with each other [29]. In [43], Averitt et al. proposed a GAN-based model to learn feature-balancing weight and to support unbiased causal estimation in the absence of unobserved confounding. Bica et al., [44] introduced a counterfactual recurrent network that leverages the increasingly available patient observational data to estimate treatment effects over time and answer such medical questions. In this study, we tackled the multi-center clinical dataset adaptation problem by using the adversarial learning framework. Instead of using private encoders to extract center-specific features, our model is capable of capturing center specificity via an external KG, which makes our model flexible to be adapted to the setup of shared-private encoders.

Notably, an external KG represents an abstraction of the real world, focusing on concepts and the interactions between the concepts [45,46]. It has been proved that KG has great power in not only representing and storing essential information about the real world, but also providing a useful tool to utilize the knowledge for potential applications. In the

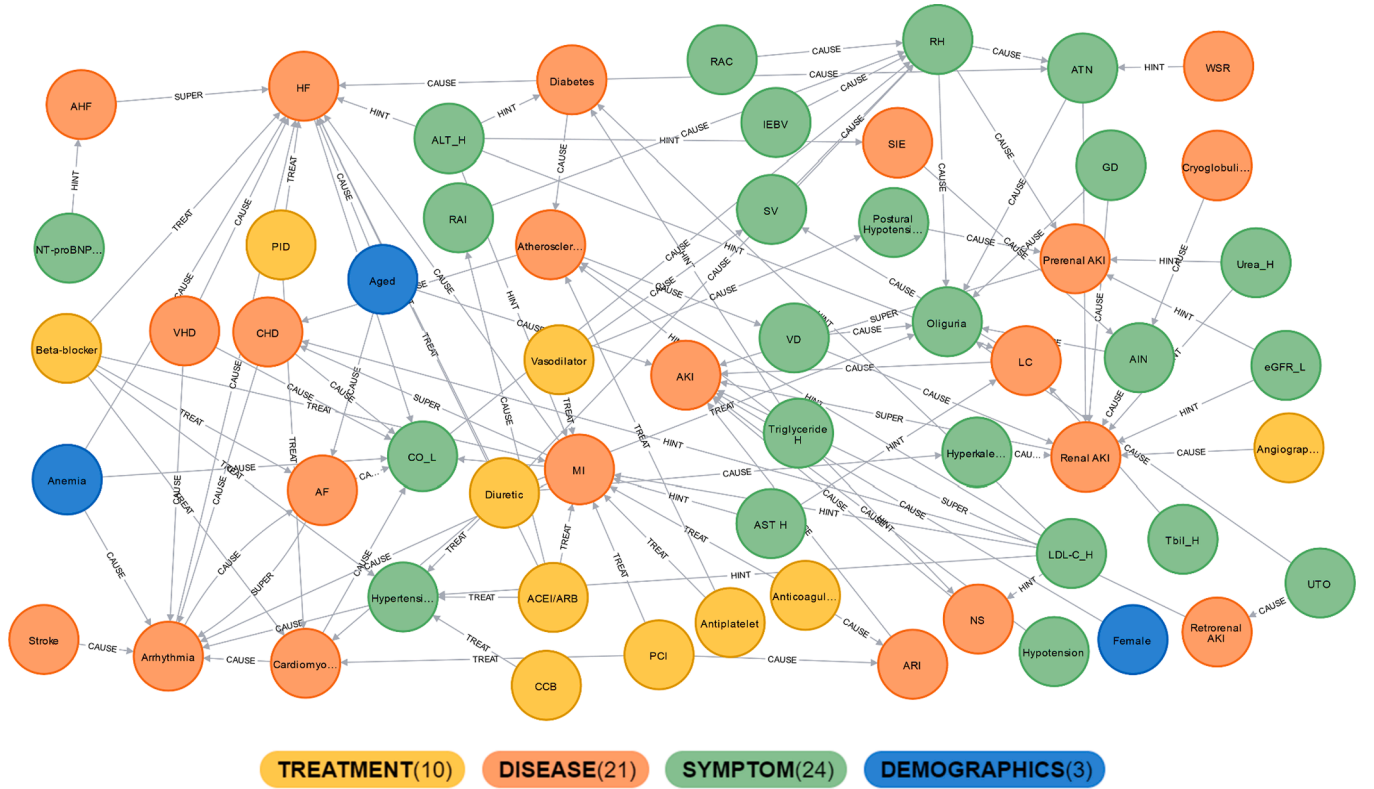


Fig. 2. Part of knowledge graph used for the task of AKI prediction in patients with HF. “H” and “L” indicate a higher- and lower-value, respectively, than the normal range of the corresponding feature.

clinical domain, KG provides the ability to search essential information much more efficiently, and to use the retrieved information for clinical purposes to improve patient care, and revolutionize care models and medical research. With the advent of graph neural networks (GNN), learning from KG has become the new trend in healthcare for diverse medical applications such as prognosis, prediction, and treatment effect estimation. For instance, Bakal et al. developed a supervised model for predicting the treatment and causative relations between biomedical entities based only on semantic graph pattern features extracted from biomedical KGs [47]. Choi et al. proposed a KG-based attention model (GRAM) that supplements electronic health records (EHR) with hierarchical information inherent to medical ontology [48]. Regarding the addressed problem in this study, KG-based multi-center clinical dataset adaptation is categorized based on the availability of cross-center connected KG, on which co-regularized training and joint embedding can be applied for learning shared representations of data. Notably, we share a similar motivation to previous studies by utilizing GNN to learn clinical concept representation in our cross-center KG. However, instead of using explicit divergence measures or adversarial losses for clinical center invariance, we present a knowledge-based attention mechanism to fully exploit the interactions between the extracted knowledge features and raw patient features to model the underlying shared information.

3. Problem definition

Multi-center clinical dataset adaptation enables the training of learning models that can perform reliable inference on heterogeneous data collected from multiple clinical centers with different settings. It assumes that both the patient features and labels across multiple clinical datasets shares the same semantic concepts but with significant discrepancies in their data distribution. In our problem setup, we considered a source clinical dataset D_s and target dataset D_t , collected from two

different clinical centers and with different marginal and conditioned distributions, i.e., $P_{D_s}(x) \neq P_{D_t}(x)$ and $P_{D_s}(x|y) \neq P_{D_t}(x|y)$. This problem, also called as the covariate shift [49], is predominant in various clinical settings, and arises primarily owing to different patient physical conditions followed with different clinical workflows at the different clinical centers. The shift may results in difference in their associations with the clinical outcomes, and thus inevitably degenerate the performance of potential clinical applications, such as outcome prediction addressed in this study.

We addressed the problem of multi-center clinical dataset adaptation in an unsupervised manner. Specifically, we assumed a clinical scenario where a prediction model is built on both the source clinical dataset with labeled patient samples $D_s^l = \{(x_i, y_i)\}_{i=1}^{N_s}$ and the target dataset with unlabeled samples $D_t^u = \{(x_i)\}_{i=1}^{N_t}$. This assumption is realistic as creating annotations for clinical data is usually expensive and time consuming. Given this problem setup, the objectives of this study were (1) to capture center-general and -specific patient representations from both source and target datasets, and (2) to efficiently predict clinical outcomes on the target dataset via the learned patient representations.

4. Method

In this section, we present our Knowledge-Aware Multi-center clinical dataset Adaptation model (KAMA), which comprises four modules: knowledge feature representation, knowledge-aware patient feature representation, adversarial learning for multi-center discrimination, and clinical outcome prediction, as shown in Fig. 1. We illustrate the details of our proposed model in this section.

4.1. Knowledge feature representation

We incorporated an external KG into our model. This KG contains clinical center-specific medical knowledge that benefits the adaptation

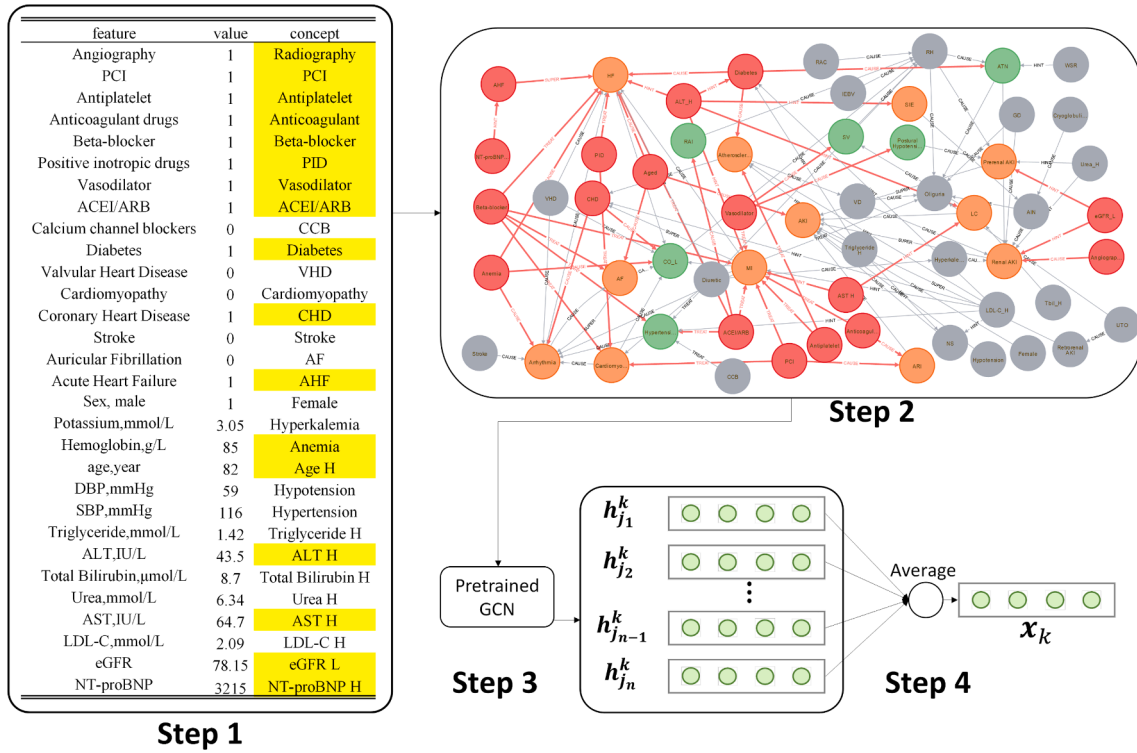


Fig. 3. The knowledge feature extraction process. (1) Step 1: we extract a set of concepts \mathcal{E}_x corresponding to the covariates in the patient sample x . The concepts that should be included in \mathcal{E}_x for this example are marked in yellow; (2) Step 2: we extract the subgraph \mathcal{G}_v . The node of the concepts extracted in step 1 are marked in red; The nodes within the vicinity of one step forward pass are marked in their original color; The rest of nodes are marked in grey; (3) Step 3: We use the pre-trained GCN model to obtain the knowledge feature vector for all nodes in the subgraph; (4) Step 4: we average over the feature vectors h_j^k for all nodes in \mathcal{G}_v to obtain the knowledge feature representation x^k for patient sample x . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

across multi-center clinical datasets as well as the improvement of performance of further applications.

We first built a KG related to a specific disease or clinical problem, such as AKI prediction for patients with HF, which is addressed in this study. To this end, we considered clinical literature, protocols, and guidelines of the addressed experimental problem as the source for constructing the corresponding KG. Formally, we defined a KG $\mathcal{G} = (\mathcal{E}, \mathcal{R})$, where \mathcal{E} and \mathcal{R} are the set of concepts and relations in the KG, respectively. In our task, a patient covariate $v \in V$ corresponds to a clinical concept $e \in \mathcal{E}$, and the relation between the concepts corresponds to a relation $r \in \mathcal{R}$. The relation set \mathcal{R} contains the following four types:

- Super-subordinate relation, where a concept is a sub-class of another, e.g., acute heart failure and chronic heart failure are both the sub-classes of heart failure.
- Treat relation, where one concept can act as the treatment intervention of another that represents a medical problem or disease; for example, beta-blocker is a common treatment of myocardial infarction.
- Cause relation, where one concept can be a risk factor of another that represents a medical problem or disease; for example, diabetes is a potential risk factor of heart failure.
- Hint relation, where one concept is an indicator of another; for example, increased serum sodium indicates water-sodium retention.

The KG built for this study contains 69 nodes of concepts, and 130 connections of relations between the nodes. Fig. 2 shows a part of the KG used in this study. The whole KG file, together with the implementation code of our model, is available on <https://github.com/ZJU-BMI/KAMA> (Fig. 3.).

After the KG was constructed, it was required to learn the KG embedding. Specifically, for each $e_i \in \mathcal{E}$, its corresponding feature vector g_{e_i} was initialized randomly and thereafter transformed into the aggregated feature vector $h_{e_i}^{(l)} \in \mathbb{R}^{d_e}$ using a graph convolution process as follows:

$$h_{e_i}^{(1)} = f \left(\sum_{r \in \mathcal{R}} \sum_{j \in N_i^r} \frac{1}{|N_i^r|} W_r^{(1)} g_{e_j} + W_0^{(1)} g_{e_i} \right) \quad (1)$$

$$h_{e_i}^{(l)} = f \left(\sum_{r \in \mathcal{R}} \sum_{j \in N_i^r} \frac{1}{|N_i^r|} W_r^{(l)} h_{e_j}^{(l-1)} + W_0^{(l)} h_{e_i}^{(l-1)} \right) \quad (2)$$

where N_i^r is the neighboring nodes of concept e_i under relation r , $f(\cdot)$ is an activation function such as ReLU, $W_r^{(l)}$ and $W_0^{(l)}$ are learnable parameters of the transformation, and l is the number of layers in the GCN.

Notably, the stack of transformations helps accumulate the normalized sum of the local neighborhood information for each concept in KG. After the clinical concept encoding, we assigned a particular score $s(e_i, r, e_j)$ to a possible triple (e_i, r, e_j) , $e_i, e_j \in \mathcal{E}$ and $r \in \mathcal{R}$. Inspired by [50], we employed DistMult factorization as the scoring function $s: \mathcal{E} \times \mathcal{R} \times \mathcal{E} \rightarrow \mathbb{R}$, which is defined as follows:

$$s(e_i, r, e_j) = \sigma(h_{e_i}^\top R_r h_{e_j}) \quad (3)$$

where σ is the logistic function, and $h_{e_i}, h_{e_j} \in \mathbb{R}^{d_e}$ are the encoded feature vectors for clinical concepts e_i and e_j . Each relation r is also associated with a diagonal matrix $R_r \in \mathbb{R}^{d_e \times d_e}$.

Thereafter, as suggested in the literature on KG representation learning [51–53], we trained the constructed KG such that for each

correct triple $(v) \in \mathcal{H}$ and incorrect triple $(e_i, r, e_j) \notin \mathcal{H}$, the training model assigns scores $s(e_i, r, e_j) > s(e_i, r, e_j')$. The task is set as a binary classification between the correct and incorrect triplets with the standard cross-entropy loss:

$$\mathcal{L}_{\mathcal{G}} = -\frac{1}{2|\mathcal{T}|} \sum_{((e_i, r, e_j), y) \in \mathcal{T}} (y \log s(e_i, r, e_j) + (1 - y) \log (1 - s(e_i, r, e_j))) \quad (4)$$

Once the KG embedding is learned, the covariate information of the particular patient sample $x_{s/t}$ contained therein can be captured in the form of knowledge feature representation $\mathbf{h}_{x_{s/t}}^k$, using the encoded concept representations. This will be effective for the downstream task when there is a distributional shift between source and target datasets during the model learning. The knowledge feature extraction process is as follows:

- (1) We extract a set of concepts \mathcal{E}_x corresponding to the covariates in the patient sample x . Notably, only the concepts with the conforming values of the corresponding features are included in \mathcal{E}_x . For example, the upper limit of normal range of alanine aminotransferase (ALT) is 40 IU/L, 43.5 IU/L indicating that ALT is higher than normal, thus the corresponding concept "ALT_H" is included in \mathcal{E}_x . To give another example, the concept "CCB" means the usage of calcium channel blockers (CCB), if CCB is not used for the patient, the concept "CCB" would not be included in \mathcal{E}_x .
- (2) Then, we extract a subgraph \mathcal{G}_V from \mathcal{G} , where we consider all the nodes either corresponding to the concepts in \mathcal{E}_x or within the vicinity of one step forward pass of concepts in \mathcal{E}_x .
- (3) We use the pre-trained GCN model to obtain the knowledge feature vector \mathbf{h}_j^k for all unique nodes j in \mathcal{G}_V .
- (4) Finally, we average over the feature vectors \mathbf{h}_j^k for all nodes in \mathcal{G}_V to obtain the knowledge feature representation \mathbf{x}^k for patient sample x , which inherently captures the center-general and -specific information likely to be beneficial to clinical dataset adaptation.

4.2. Knowledge-aware patient feature representation

We input both the original patient sample \mathbf{x}^p and the extracted knowledge feature representation \mathbf{x}^k into the patient feature encoder and knowledge feature encoder, to extract the latent patient feature representation \mathbf{z}^p and latent knowledge feature representation \mathbf{z}^k , respectively.

To further enforce the invariant knowledge feature representation \mathbf{x}^k , we considered it as a latent representation in a traditional autoencoder and thereafter equipped the proposed model with an extra decoder D_{rec} with parameters Θ_R to calculate the reconstruction loss between the input \mathbf{x}^k and the reconstructed feature vector $\hat{\mathbf{x}}^k$:

$$\mathcal{L}_{rec}(\mathbf{x}^k) = -\mathbb{E}_{\mathbf{x}^k} \left(\|D_{rec}(\mathbf{z}^k) - \mathbf{x}^k\|_2^2 \right) \quad (5)$$

Notably, D_{rec} is used to reconstruct the knowledge feature representation for both the source and target patient samples such that clinical center-invariant features \mathbf{z}^k can be extracted.

We argue that the interactive relation between \mathbf{z}^p and \mathbf{z}^k can provide clues to focus on the salient features for potential clinical applications. To this end, we designed a hierarchical attention mechanism, including feature-level attention and sample-level attention, to explore the semantic compositionality between \mathbf{z}^p and \mathbf{z}^k .

We developed a knowledge-aware mutual attention mechanism for capturing the correlation between patient feature and knowledge feature representations. Formally, we used the dot product between \mathbf{z}^p

and \mathbf{z}^k to measure the correlation matrix \mathbf{M} for the input patient sample, given as:

$$\mathbf{M} = (\mathbf{z}^p)^\top \cdot \mathbf{z}^k \quad (6)$$

where each element in \mathbf{M}_{ij} represents the correlation signal between the i th element in the patient feature representation \mathbf{z}^p and the j th element in the knowledge feature representation \mathbf{z}^k . It should be noted that we hypothesized that patient covariates have different contributions to the overall patient feature representation, and the feature-level attention mechanism can identify essential patient features to represent the patient clinical condition. To this end, we measured the normalized weight vectors α^p and α^k of each row and column of \mathbf{M} as attention signals through the SoftMax function:

$$\alpha^p = \text{SoftMax} \left(\frac{\sum_{i=1}^N \mathbf{M}[:, i]}{N} \right) \quad (7)$$

$$\alpha^k = \text{SoftMax} \left(\frac{\sum_{j=1}^N \mathbf{M}[j, \cdot]}{N} \right) \quad (8)$$

Thereafter, we extracted the knowledge-aware patient feature representation matrix \mathbf{B}^p and the patient-oriented knowledge feature representation matrix \mathbf{B}^k as follows:

$$\mathbf{B}^p = \tanh(\mathbf{U}_{p1}(\mathbf{z}^p + (\mathbf{I}^p \otimes \alpha^p) \odot \mathbf{z}^k)) \quad (9)$$

$$\mathbf{B}^k = \tanh(\mathbf{U}_{k1}(\mathbf{z}^k + (\mathbf{I}^k \otimes \alpha^k) \odot \mathbf{z}^p)) \quad (10)$$

where \mathbf{U}_{p1} and \mathbf{U}_{k1} are projection parameters, $\mathbf{I}^p, \mathbf{I}^k = [1, \dots, 1]^\top$ indicates a N dimensional all-ones vector, \otimes denotes the Kronecker product operation, and \odot denotes the element-wise multiplication.

Based on \mathbf{B}^p and \mathbf{B}^k , the final knowledge-aware patient feature representation π^p and the patient-oriented knowledge feature representation π^k can be measured as follows:

$$\pi^p = \mathbf{B}^p \cdot \mathbf{z}^p \quad (11)$$

$$\pi^k = \mathbf{B}^k \cdot \mathbf{z}^k \quad (12)$$

Both π^p and π^k were then concatenated to form the final knowledge-aware patient representation $\pi(\mathbf{x}) = [\pi^p; \pi^k]$ for the input patient sample \mathbf{x} .

4.3. Adversarial learning for multi-center discrimination

We employed an adversarial learning module to learn a clinical center-invariant mapping between source and target patient samples. Specifically, the knowledge-aware module is considered as the generator in adversarial learning. As can be seen in Fig. 1(b), the generator was used to generate the latent representations with the samples from the source and the target datasets, denoted as $\pi(\mathbf{x}_s) = [\pi_s^p; \pi_s^k]$ and $\pi(\mathbf{x}_t) = [\pi_t^p; \pi_t^k]$, respectively. Notably, \mathbf{x}_s and \mathbf{x}_t represents the samples from the source and the target datasets, respectively, and they are not paired or related. Meanwhile, as shown in Fig. 1(a), a multi-center discriminator \mathcal{D} with parameters $\Theta_{\mathcal{D}}$ was used to distinguish between latent representations of the source and target patient samples $\pi(\mathbf{x}_s)$ and $\pi(\mathbf{x}_t)$. $\mathcal{D}(\pi(\mathbf{x}))$ is the probability predicted by \mathcal{D} , that the representation $\pi(\mathbf{x})$ is from the source dataset. The loss of adversarial learning is defined as the cross-entropy loss:

$$\mathcal{L}_{adv} = -\mathbb{E}_{\mathbf{x}_s} (\log \mathcal{D}(\pi(\mathbf{x}_s))) - \mathbb{E}_{\mathbf{x}_t} (\log (1 - \mathcal{D}(\pi(\mathbf{x}_t)))) \quad (13)$$

The generator and the discriminator are optimized adversarially and alternately by contesting with each other, i.e., the discriminator is trained to distinguish the latent representations of the source and target patient samples, while the generator aims to confuse the discriminator. In the training process, the discriminator forces the generator to extract

Table 1

Baseline comparison between the PLAGH and MIMIC datasets.

Feature	MIMIC Dataset (1,006)	PLAGH Dataset (5,075)	p-value
AKI	0.451 (454)	0.071 (365)	<0.05
In-hospital mortality	0.097 (98)	0.027 (139)	<0.05
ACEI/ARB	0.45 (451)	0.50 (2547)	<0.05
Acute HF	0.25 (247)	0.34 (1723)	<0.05
AF	0.33 (331)	0.22 (1121)	<0.05
Age, year	66.52 (56.57–77.19)	61.00 (51.00–70.00)	<0.05
ALT, IU/L	28.00 (18.00–50.00)	21.00 (14.40–33.80)	<0.05
Angiography	0.49 (496)	0.40 (2008)	<0.05
Anticoagulant	0.77 (773)	0.38 (1927)	<0.05
Antiplatelet	0.64 (642)	0.65 (3298)	0.501
AST, IU/L	35.00 (22.00–68.89)	21.30 (16.30–30.45)	<0.05
Beta-blocker	0.70 (707)	0.68 (3428)	0.097
Cardiomyopathy	0.12 (122)	0.19 (941)	<0.05
CCB	0.15 (154)	0.22 (1110)	<0.05
CHD	0.68 (682)	0.58 (2928)	<0.05
Diabetes	0.26 (260)	0.39 (2002)	<0.05
DBP, mmHg	64.50 (55.00–74.00)	74.00 (67.00–81.00)	<0.05
eGFR, mL/min/1.73 m ²	83.89 (71.50–94.88)	87.63 (75.66–98.80)	<0.05
Hemoglobin, g/L	125.00 (110.00–140.00)	137.00 (124.00–150.00)	<0.05
LDL-C, mmol/L	2.02 (1.62–2.53)	2.25 (1.79–2.80)	<0.05
NT-pro-BNP, pg/mL	2140.29 (1170.51–4160.60)	1215.84 (422.59–2949.00)	<0.05
PCI	0.20 (205)	0.19 (969)	0.369
PID	0.19 (191)	0.37 (1867)	<0.05
Potassium, mmol/L	4.10 (3.80–4.40)	3.89 (3.62–4.17)	<0.05
SBP, mmHg	121.00 (106.00–136.00)	124.80 (113.00–138.00)	<0.05
Sex, male	0.59 (597)	0.68 (3431)	<0.05
Stroke	0.06 (59)	0.10 (485)	<0.05
Urea, mmol/L	6.07 (4.64–7.85)	5.84 (4.73–7.26)	<0.05
Total bilirubin, μmol/L	10.26 (6.84–15.39)	13.70 (9.80–19.90)	<0.05
Triglyceride, mmol/L	1.28 (1.02–1.62)	1.12 (0.83–1.59)	<0.05
Vasodilator	0.30 (302)	0.61 (3103)	<0.05
VHD	0.25 (249)	0.12 (616)	<0.05

Abbreviations: ACEI: angiotensin-converting enzyme inhibitors; ALT, alanine aminotransferase; ARB, angiotensin receptor blockers; AST, aspartic-transaminase; CCB, calcium channel blockers; CHD, coronary heart disease; DBP, diastolic blood pressure; eGFR, estimated glomerular filtration rate; LDL-C, low density lipoprotein cholesterol; NT-pro-BNP, N-terminal pro-brain natriuretic peptide; PCI, percutaneous coronary intervention; PID, positive inotropic drugs; SBP, systemic blood pressure; TBil, total bilirubin; VHD, valvular heart disease.

* most NT-pro-BNP values in MIMIC dataset were imputed.

the latent representations containing shared information in both datasets, which is unidentifiable for the discriminator. After training, the similarity between the distributions of different datasets can be captured by the generator and the information is encoded into the representations, which finally benefits the dataset adaptation.

4.4. Clinical outcome prediction

By learning the latent shared patient representations $\pi(x)$ for both the source and target patient samples, a clinical outcome predictor \mathcal{C} with parameters $\Theta_{\mathcal{C}}$ trained for the source dataset can be smoothly applied for the target patient samples.

As can be seen in Fig. 1(a), the clinical outcome predictor \mathcal{C} takes the latent shared patient representations $\pi(x)$ as the input and output the probability \hat{y} that the label is positive, denoted as $\hat{y} = \mathcal{C}(\pi(x))$. \hat{y}_s and \hat{y}_t denote the predicted labels with the input features from the source and the target dataset, respectively. Specifically, \mathcal{C} consists of fully-connected layers, and the activation function of the last layer is sigmoid function for the purpose of outputting the probability within the

range (0,1). The loss function of clinical outcome prediction is defined as the cross-entropy loss:

$$\mathcal{L}_{cls} = -\mathbb{E}_{(x_s, y_s)} (y_s \log(\hat{y}_s) + (1 - y_s) \log(1 - \hat{y}_s)) \quad (14)$$

The final optimization of the proposed model is based on the min-max objective as follows:

$$\Theta^* = \arg \min_{\Theta_{\pi}, \Theta_R, \Theta_{\mathcal{C}}} \max_{\Theta_{\mathcal{D}}} (\mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{rec}) \quad (15)$$

where λ_1 and λ_2 are hyper-parameters.

5. Experiments

5.1. Datasets

To evaluate our proposed model, we experimented on two real clinical datasets that include patients admitted for HF with normal renal function. One dataset was collected from the Chinese PLA General hospital (PLAGH) and includes 5075 patients; the other is a public dataset extracted from MIMIC III and includes 1006 patients. As the PLAGH dataset contains patient data mainly from the general wards in PLAGH, and the MIMIC dataset that from ICUs in the United States, there are inevitably differences between the baseline characteristics of the patients in the two datasets. As can be seen in Tables 1 and 2, there are statistically significant differences in most features between the whole datasets as well as the subgroups conditioned on AKI occurrence, i.e., $P_{PLAGH}(x) \neq P_{MIMIC}(x)$ and $P_{PLAGH}(x|y) \neq P_{MIMIC}(x|y)$, indicating the existence of dataset shift. Notably, to calculate the p-values, Mann-Whitney U test was used for the features with continuous values, while Chi-squared test was used for the features with binary values. All the statistical tests are two-sided.

In this sense, both multi-center clinical datasets can be used to evaluate the stability of learning an effective AKI prediction model for patients with HF in diverse clinical settings.

5.2. Experimental configurations

Our model consists of a patient feature encoder, a knowledge feature encoder, a knowledge feature decoder, a clinical center-discriminator, and a clinical outcome predictor. We implemented each of them with three fully connected layers with 20% dropout rate. For initialization, all parameters were randomly initialized. The initial learning rate α was set to 0.001. The loss weight coefficient λ_1 was set to 0.5, and λ_2 was set to 1.0. The model was trained with Adam optimizer [54].

5.3. Baseline models

We compared our KAMA model with several state-of-the-art baselines, which are described as follows:

- (1) Multi-layer perceptron (MLP): We trained an MLP-based prediction model on the source dataset and then evaluated it on the target dataset.
- (2) MLP + adversarial learning (MLP + AL): We used an adversarial neural network for clinical center-discriminator so that center-invariant features could be captured. Then, we used the MLP for the clinical outcome prediction.
- (3) Central moment discrepancy (CMD) [40]: This is a regularization method to minimize the difference between feature representations by utilizing equivalent representation of probability distribution on both source and target datasets.
- (4) Support vector machine (SVM): We used an SVM-based prediction model on the source dataset and then evaluated it on the target dataset.
- (5) Transductive SVM (TSVM) [55,56]: This is a semi-supervised learning algorithm that uses the unlabeled samples while

Table 2

Baseline comparison between the PLAGH and MIMIC datasets conditioned on AKI occurrence.

Feature	AKI = 0			AKI = 1		
	MIMIC Dataset (552)	PLAGH Dataset (4,710)	p-value	MIMIC Dataset (454)	PLAGH Dataset (365)	p-value
ACEI/ARB	0.49 (273)	0.51 (2410)	0.474	0.39 (178)	0.38 (137)	0.677
Acute HF	0.24 (134)	0.33 (1561)	<0.05	0.25 (113)	0.44 (162)	<0.05
AF	0.28 (156)	0.22 (1015)	<0.05	0.39 (175)	0.29(106)	<0.05
Age, year	65.97 (56.16–77.11)	60.00 (50.00–70.00)	<0.05	67.41 (57.05–77.29)	68.00 (58.00–77.00)	0.780
ALT, IU/L	29.00 (18.00–54.00)	21.00 (14.50–33.60)	<0.05	27.00 (18.00–46.00)	22.20 (13.10–38.40)	<0.05
Angiography	0.50 (274)	0.40 (1867)	<0.05	0.49 (222)	0.39 (141)	<0.05
Anticoagulant	0.79 (434)	0.36 (1719)	<0.05	0.75 (339)	0.57 (208)	<0.05
Antiplatelet	0.68 (376)	0.65 (3076)	0.205	0.59 (266)	0.61 (222)	0.565
AST, IU/L	35.50 (23.00–68.66)	21.10 (16.20–29.70)	<0.05	34.00 (22.00–68.75)	25.30 (18.70–49.30)	<0.01
Beta-blocker	0.73 (403)	0.68 (3215)	<0.05	0.67 (304)	0.58 (213)	<0.05
Cardiomyopathy	0.13 (72)	0.19 (911)	<0.05	0.11 (50)	0.08 (30)	0.222
CCB	0.13 (73)	0.22 (1036)	<0.05	0.18 (81)	0.20 (74)	0.427
CHD	0.68 (373)	0.57 (2686)	<0.05	0.68 (309)	0.66 (242)	0.646
DBP, mmHg	66.00 (57.00–76.00)	74.00 (68.00–81.00)	<0.05	62.00 (53.00–73.00)	72.38 (65.32–80.00)	<0.05
Diabetes	0.25 (139)	0.38 (1799)	<0.05	0.27 (121)	0.56 (203)	<0.05
eGFR, mL/min/1.73 m ²	83.30 (70.97–93.83)	87.87 (75.89–98.89)	<0.05	85.07 (71.89–96.49)	84.66 (71.93–97.36)	0.624
Hemoglobin, g/L	128.00 (113.00–143.00)	138.00 (125.00–150.00)	<0.05	121.00 (107.00–135.00)	127.00 (113.00–143.00)	<0.05
LDL-C, mmol/L	2.08 (1.67–2.58)	2.25 (1.79–2.81)	<0.05	1.96 (1.57–2.46)	2.22 (1.72–2.80)	<0.05
NT-pro-BNP, pg/mL	1926.31 (1119.19–3965.50)	1153.05 (398.34–2710.00)	<0.05	2277.84 (1251.48–4230.59)	2880.00 (1134.00–6081.00)	<0.05
PCI	0.25 (138)	0.19 (893)	<0.05	0.15 (67)	0.21 (76)	<0.05
PID	0.18 (101)	0.36 (1699)	<0.05	0.20 (90)	0.46 (168)	<0.05
Potassium, mmol/L	4.10 (3.80–4.40)	3.89 (3.62–4.16)	<0.05	4.10 (3.80–4.40)	3.87 (3.53–4.24)	<0.05
SBP, mmHg	122.00 (108.00–136.00)	124.87 (113.00–138.00)	<0.05	119.00 (104.00–137.00)	124.00 (111.63–138.00)	<0.05
Sex, male	0.62 (344)	0.68 (3202)	<0.05	0.56 (253)	0.63 (229)	0.051
Stroke	0.06 (33)	0.09 (436)	<0.05	0.06 (26)	0.13 (49)	<0.05
Total bilirubin, μ mol/L	10.26 (6.84–15.39)	13.70 (9.80–19.70)	<0.05	10.26 (6.84–15.39)	14.60 (10.10–21.60)	<0.05
Triglyceride, mmol/L	1.27 (1.03–1.63)	1.13 (0.84–1.60)	<0.05	1.29 (1.02–1.61)	1.01 (0.74–1.39)	<0.05
Urea, mmol/L	6.07 (4.64–7.85)	5.81 (4.72–7.19)	<0.05	6.43 (5.00–8.21)	6.40 (5.04–8.37)	0.302
Vasodilator	0.28 (156)	0.61 (2864)	<0.05	0.32 (146)	0.65 (239)	<0.05
VHD	0.23 (126)	0.12 (549)	<0.05	0.27 (123)	0.18 (67)	<0.05

Abbreviations: ACEI: angiotensin-converting enzyme inhibitors; ALT, alanine aminotransferase; ARB, angiotensin receptor blockers; AST, aspartic transaminase; CCB, calcium channel blockers; CHD, coronary heart disease; DBP, diastolic blood pressure; eGFR, estimated glomerular filtration rate; LDL-C, low density lipoprotein cholesterol; NT-pro-BNP, N-terminal pro-brain natriuretic peptide; PCI, percutaneous coronary intervention; PID, positive inotropic drugs; SBP, systemic blood pressure; TBil, total bilirubin; VHD, valvular heart disease.

*most NT-pro-BNP values in MIMIC dataset were imputed.

searching the separating hyperplanes of SVM classifiers. We applied it to build a clinical outcome predictor by using labeled and unlabeled data.

- (6) Semi-supervised variational autoencoder (SemiVAE) [57]: This is a semi-supervised deep generative model extended on the basis of the variational autoencoder to perform semi-supervised learning by recognizing it as a specialized missing data imputation task for the classification problem.

It should be noted that although both TSVM and SemiVAE fall outside the regime of unsupervised clinical dataset adaptation, we have reported their results to emphasize the importance of clinical dataset adaptation.

For evaluation, we used area under the receiver operating characteristic curve (AUC) to measure the performance and all the models were validated with 5-fold cross-validation. Specifically, in each fold, we trained the models on 80% samples of the source dataset, and then evaluated the models on the remaining 20% of the source dataset and the whole target dataset, respectively.

To further illustrate the ability of our model on data adaptation, we conducted experiments with the following two additional models:

- (1) MLP with data after PSM (MLP-PSM): We used PSM to pick out the matched samples in both datasets. We trained the MLP on the matched part of the source dataset and then evaluated it on its unmatched part (source-unmatched), the matched part of the target dataset (target-matched), and the whole target dataset (target-whole).
- (2) MLP with simply merged labeled datasets (MLP-MERGE): We setup a best-case scenario where the labels of both the datasets

were available. We added the dataset identifier as an additional input variable and then merged the two datasets. We used MLP to conduct the prediction, and evaluated the performance with 5-fold cross-validation.

5.4. Results and analysis

We compared the performance of the proposed model with the baselines. As observed in Fig. 4, MLP + AL outperforms MLP by 5.0% and 6.3%, on the target datasets with respect to the tasks PLAGH→MIMIC and MIMIC→PLAGH, respectively. This result indicates that equipping the model with the adversarial learning strategy can boost the prediction performance on the target datasets. Similarly, KAMA outperforms MLP by 7.0% and 8.4% on the target datasets for the respective tasks. This observation demonstrates that a significant distribution bias exists between the source and target datasets, which degrades the prediction performance of applying the learned model from the source to the target dataset. However, the center-invariant features can alleviate this problem. Comparatively, KAMA surpasses MLP + AL by 2.0% and 2.1% on the target dataset for both tasks PLAGH→MIMIC and MIMIC→PLAGH, respectively, indicating the remarkable improvement in clinical center invariance owing the incorporation of an external KG.

Table 3 shows the results achieved on the datasets in comparison with baselines. As indicated in [49], transferring learning across two datasets with a significant distribution bias is quite challenging. CMD is a state-of-the-art method of transfer learning, and our model outperforms it by approximately 2.4% and 3.5% overall on the target dataset for the two tasks, respectively. As observed, our proposed KAMA achieved remarkable performance gains in both the task scenarios, indicating that the incorporation of external knowledge can help in

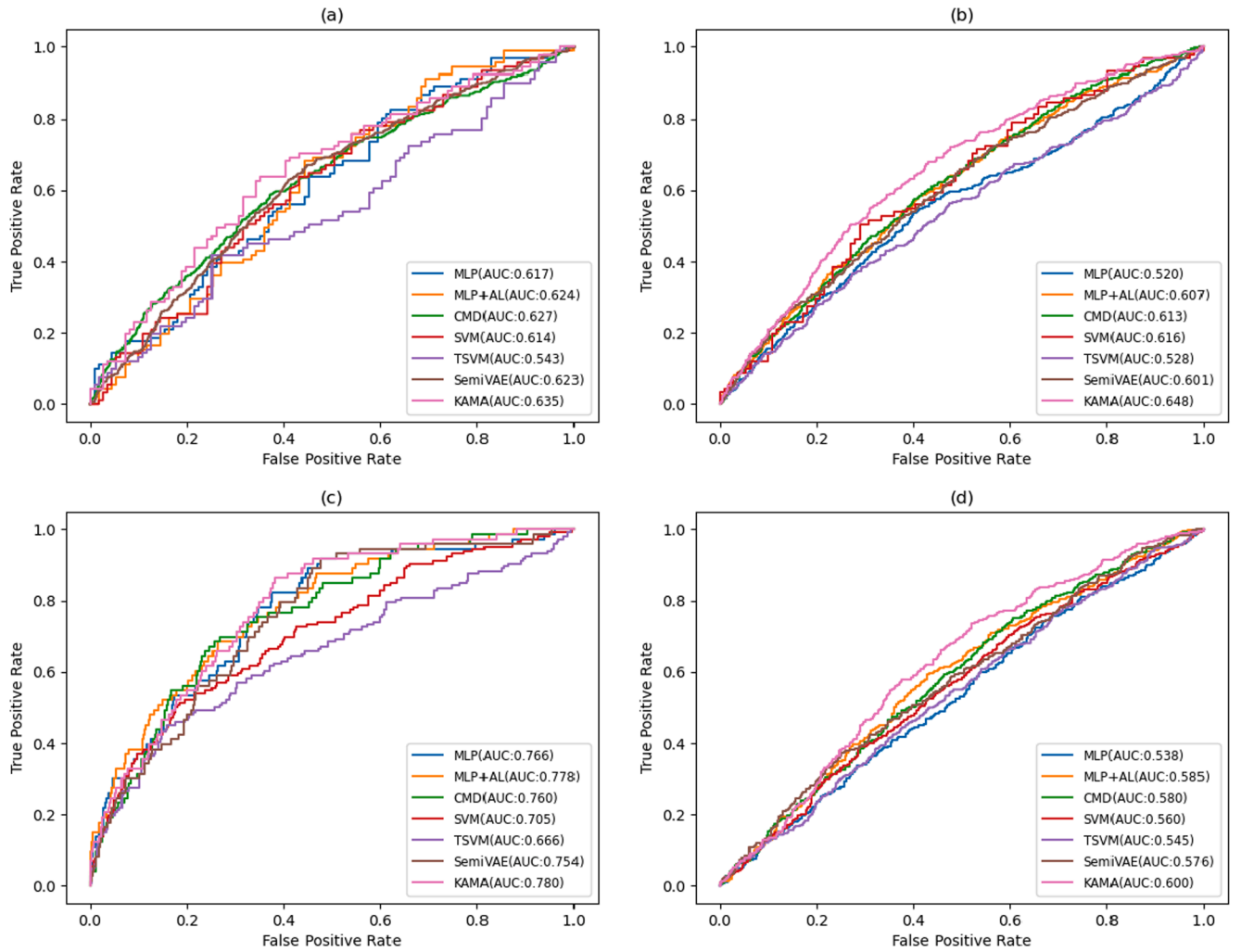


Fig. 4. Results of the proposed model and baselines across multi-center clinical datasets. Receiver operating characteristic (ROC) curves of models achieved on (a) the source MIMIC dataset and (b) the target PLAGH dataset for the task MIMIC→PLAGH. ROC curves of models achieved on (c) the source PLAGH dataset and (d) the target MIMIC dataset for the task PLAGH→MIMIC.

Table 3

Comparison of the proposed model with different baseline and state-of-the-art models.

	PLAGH→MIMIC		MIMIC→PLAGH			
	source_AUC (PLAGH)	target_AUC (MIMIC)	source_AUC (MIMIC)	target_AUC (PLAGH)		
MLP	0.766 ± 0.007	0.538 ± 0.009	0.617 ± 0.005	0.520 ± 0.040		
MLP + AL	0.778 ± 0.002	0.585 ± 0.002	0.624 ± 0.005	0.607 ± 0.010		
CMD	0.760 ± 0.009	0.580 ± 0.003	0.627 ± 0.006	0.613 ± 0.004		
SVM	0.705 ± 0.000	0.560 ± 0.000	0.614 ± 0.001	0.616 ± 0.011		
TSVM	0.666 ± 0.000	0.545 ± 0.000	0.543 ± 0.000	0.528 ± 0.000		
SemiVAE	0.754 ± 0.003	0.576 ± 0.004	0.623 ± 0.004	0.601 ± 0.013		
KAMA	0.780 ± 0.002	0.600 ± 0.002	0.635 ± 0.002	0.648 ± 0.006		
	Source-unmatched	Target-matched	Target-whole	Source-unmatched	Target-matched	Target-whole
MLP-PSM (0.02)*	0.527±0.012	0.723±0.039	0.544±0.009	0.544±0.021	0.553±0.003	0.510±0.012
MLP-PSM (0.1)*	0.620±0.025	0.739±0.036	0.589±0.007	0.548±0.030	0.577±0.006	0.545±0.042
MLP-PSM (1)*	0.659±0.021	0.678±0.018	0.575±0.009	0.559±0.015	0.569±0.009	0.551±0.020
MLP-PSM (2)*	0.679±0.018	0.661±0.028	0.571±0.007	0.568±0.013	0.555±0.013	0.547±0.017
	PLAGH+MIMIC(MERGE)→PLAGH			PLAGH+MIMIC(MERGE)→MIMIC		
MLP-MERGE	0.792±0.002			0.656±0.007		

* The numbers in the parentheses are the match tolerances of PSM.

learning latent patient representations across multi-center clinical datasets and thereafter even surpass sophisticated state-of-the-art models on clinical outcome prediction.

In addition, the performance of TSVM and SemiVAE is worth discussion. As a semi-supervised learning method, TSVM exhibited the worst performance among the baselines owing to the significant

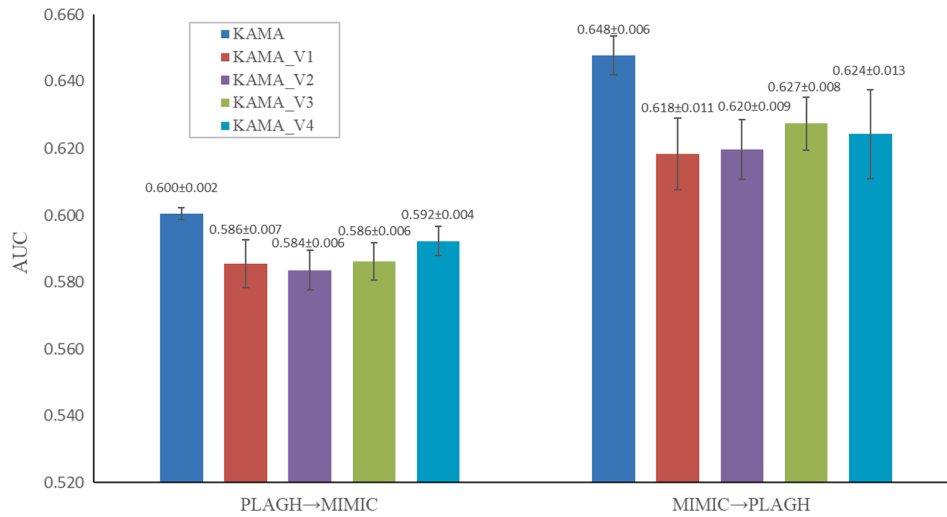


Fig. 5. AUC on target dataset across different variants of proposed model.

distribution bias between both the datasets. Meanwhile, SemiVAE achieved a relatively better performance compared to TSVM. As a representation-based semi-supervised method, SemiVAE works similarly as other representation learning methods when applied for data adaptation. However, without proper and adequate regularization on the representations, SemiVAE could not match the performance of even MLP + AL and CMD, much less that of the proposed KAMA with external knowledge from KG.

Moreover, the results obtained using MLP-PSM and MLP-MERGE on the experimental datasets are shown at the half bottom of Table 3. As for MLP-PSM, the size of the matched subgroup is actually fairly small (66 and 154 matched samples in MIMIC and PLAGH, respectively) when the match tolerance is 0.02 standard deviation, owing to the significant distributional discrepancy between the two datasets. As can be seen in Table 3, the model trained on the matched part of PLAGH performs well on the matched part of MIMIC; however, the performance decreases sharply when it is applied to the whole MIMIC and even the unmatched part of PLAGH.

Moreover, we investigate the impact of the value of the match tolerance on performance, and the obtained results are shown in the half bottom of Table 3 (the detailed experimental results are presented in the supplementary). With the increases of the value of the match tolerance, the achieved AUC on the unmatched subgroup of the source dataset increases, while AUCs on the matched subgroup of the target dataset and the whole target dataset increase at first, and then decrease with the further increases of the value of the match tolerance. Similar trends can be observed on both PLAGH→MIMIC and MIMIC→PLAGH tasks. It is considerable to see the increase of the AUC on the unmatched subgroup of the source dataset, since more samples are included in the training set. As for the initial increases of the AUC on the matched subgroup of either target dataset or whole target dataset, it is benefitted from the increases of sample size, especially for the MIMIC→PLAGH task, in which the training set is relatively small. However, with the increases of the value of the match tolerance, more samples with diverse distributions are involved, causing the deterioration of performance. Compared with MLP-PSM, the proposed KAMA exhibits an obvious superiority. This finding indicates that our model performs better than PSM-based models in terms of dataset adaptation, especially when there is less overlap between the datasets.

Furthermore, by comparing the performance of MLP-MERGE with the proposed KAMA, we observe that our proposed model reaches the close performance obtained in the best-case scenario, where the labels of both datasets were available (AUC: 0.792 for MLP-MERGE vs. 0.780 for KAMA on PLAGH; AUC: 0.656 for MLP-MERGE vs. 0.635 for KAMA on MIMIC). By incorporating the external KG, the proposed KAMA can

well utilize the unlabeled samples from the target dataset and capture the center-specific/invariant features to achieve remarkable performance without using the labels of the target dataset.

5.5. Ablation study

We further analyzed our model and challenged the design choices. Specifically, we considered four variants of our proposed model based on alternative approaches to condition the model with knowledge features. Variant 1 uses separate patient feature encoders for source and target datasets. In Variant 2, the clinical center-discriminator accepts only patient features π^p as input whereas the clinical outcome predictor C takes the concatenated features $[\pi^p; \pi^k]$. Variant 3 does not employ the knowledge-aware attention mechanism but inputs the encoded knowledge features and patient features into the outcome predictor and clinical center-discriminator directly. In Variant 4, the clinical center-discriminator accepts $[\pi^p; \pi^k]$ as input, whereas the outcome predictor only accepts π^p without knowledge features.

As can be seen in Fig. 5, all the variants perform worse than KAMA. The performance of Variant 1 decreases significantly in comparison with KAMA. This finding indicates that including a shared patient feature encoder in the proposed model can facilitate learning invariant representations to improve the performance of clinical outcome prediction. For Variant 2, removal of π^k from the clinical center-discriminator diminishes the center-invariance capabilities, and results in a stronger discriminator and worse clinical outcome predictor. For Variant 3, removal of the knowledge-aware attention mechanism degrades the performance of outcome prediction, indicating that exploring interactive correlations between patient features and knowledge features is helpful for capturing latent patient representations and further clinical applications. For Variant 4, removal of π^k from clinical outcome predictor degrades its performance. This indicates that knowledge features π^k contain task-specific information retrieved from the external KG.

5.6. Odd ratios of KG-concepts

We conducted a case study to observe the experimental results for insight. Specifically, patient features were matched to the corresponding KG-concepts to calculate their odds ratios in terms of AKI occurrence in HF patients on both PLAGH and MIMIC. The closer the value is to one, the less relevant is the concept in terms of AKI. An odds ratio larger than one indicates a positive correlation between the KG-concept and AKI occurrence, and otherwise negative. As can be seen in Fig. 6, it is easy to distinguish between center-general and -specific KG-concepts. For

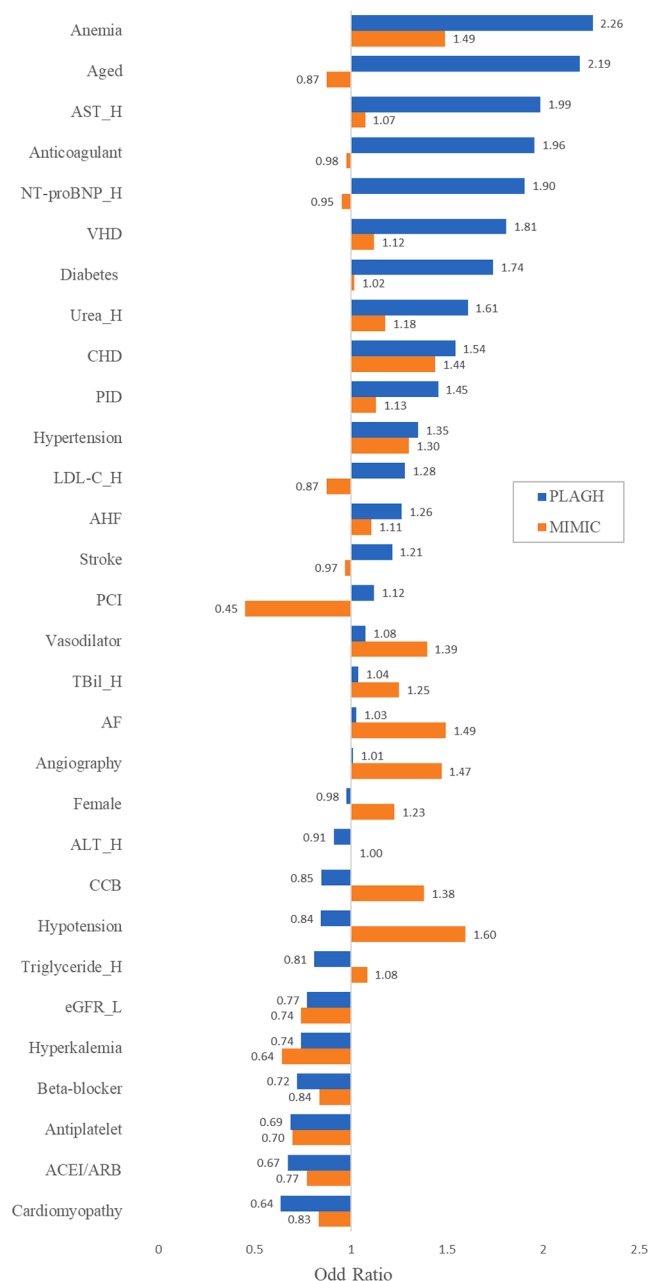


Fig. 6. Odds ratios of KG-concepts on both PLAGH and MIMIC datasets.

example, “Anemia” is a typical center-general KG-concept with odds ratios larger than one on both PLAGH and MIMIC. Clinically, anemia is one of the most common complications of HF, and can either induce HF or be caused by it [58]. The kidney of HF patients exhibit hypoperfusion, while anemia reduces the oxygen content in blood and further aggravates renal ischemia. In addition, the odds ratio of the concept “PCI” is close to one on PLAGH and less than one on MIMIC, indicating that it is ICU setting-specific. As an important treatment for patients with coronary artery stenosis, PCI can instantly improve myocardial perfusion and protect the cardiac muscle function. Cardiac muscle function ensures acceptable cardiac output and maintains the renal perfusion, which benefits the kidney. As the MIMIC dataset was collected from ICU settings, the registered patients are generally in a worse condition than the ones in PLAGH, which was collected from general wards. Considerably, PCI is more meaningful to be conducted in the MIMIC settings and is a disincentive to AKI occurrence. On the contrary, “Angiography” is also a center-specific KG-concept; however, it has an odds ratio close

to one on PLAGH and larger than one on MIMIC. Angiography is an important image examination for HF patients to check coronary artery stenosis. However, the iodic contrast agent used for angiography is a burden on the kidney. Especially for ICU patients, who are in worse condition, it is comprehensible to find angiography to be an inducement of AKI.

6. Conclusions

In this paper, we proposed a novel knowledge-aware multi-center clinical data adaptation model, which allows large-scale training based on labeled clinical data collected from the source center and unlabeled data from the target center, and the adaptation is fulfilled by aligning the distribution of features across the two datasets by employing an extra KG. We evaluated the proposed model on a case study of AKI prediction for patients with HF. The experimental results show that our proposed model is effective in extracting the common information among multi-center clinical datasets, and achieved significant improvements over state-of-the-art baselines in the prediction task.

Although the proposed model achieved impressive performance on multi-center clinical data adaptation, a few limitations still need to be overcome, and further investigation can be conducted in our future work. First, the current version of the proposed model only considers datasets collected from two medical centers. An extension is needed for the proposed model to be adaptable to datasets collected more different centers.

In addition, the primary intention of this study was to equip a deep neural network with an external KG, and to highlight the role of knowledge base infusion in multi-center clinical dataset adaptation. We use naïve neural networks simply without complex structures. Nevertheless, the proposed method of incorporating KG into neural networks is universal and flexible. The flexibility of the proposed method allows it to be associated with the more sophisticated neural networks, which we believe can make it perform even better. We intend to analyze this in future.

Moreover, we argue that the proposed model can facilitate further research into heterogeneous clinical data fusion problems. The proposed model is not confined to the clinical outcome prediction task, but can be used in several downstream clinical tasks that rely on multi-center clinical datasets. In the future, we plan to evaluate our model on a larger scale of distributed clinical data from a wide spectrum of patient samples, and exploit the potential of the proposed model, which holds a crucial advantage over traditional techniques.

Funding

This work was partially supported by the National Key Research and Development Program of China under Grant No. 2018YFC2001204 and the National Natural Science Foundation of China under Grant No. 61672450. The supporting bodies had no influence on the analysis and interpretation of data, on the writing of the report, or on the decision to submit the paper for publication.

CRediT authorship contribution statement

Jiebin Chu: Visualization, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Jinbiao Chen:** Software, Visualization. **Xiaofang Chen:** Software, Visualization. **Wei Dong:** Resources, Data curation, Supervision, Funding acquisition. **Jinlong Shi:** Resources, Data curation, Supervision, Funding acquisition. **Zhengxing Huang:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jbi.2021.103710>.

References

- [1] D. Riaño, M. Peleg, A. ten Teije, Ten years of knowledge representation for health care (2009–2018): Topics, trends, and challenges, *Artif. Intell. Med.* 100 (2019), 101713.
- [2] E. Rossi, M. Rosa, L. Rossi, A. Priori, S. Marceglia, WebBioBank: A new platform for integrating clinical forms and shared neurosignal analyses to support multi-centre studies in Parkinson's Disease, *J. Biomed. Inform.* 52 (2014) 92–104.
- [3] H. Kondylakis, B. Claerhout, M. Keyur, et al., The INTEGRATE project: Delivering solutions for efficient multi-centric clinical research and trials, *J. Biomed. Inform.* 62 (2016) 32–47.
- [4] J. Waring, C. Lindvall, R. Umeton, Automated machine learning: Review of the state-of-the-art and opportunities for healthcare, *Artif. Intell. Med.* 104 (2020), 101822.
- [5] D. Ben-Israel, W.B. Jacobs, S. Casha, et al., The impact of machine learning on patient care: A systematic review, *Artif. Intell. Med.* 103 (2020), 101785.
- [6] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M. A. Suchard, R.W. Park, I.C. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G. N. Norén, Y.C. Li, P.E. Stang, D. Madigan, P.B. Ryan, Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers, *Stud. Health Technol. Informatics* 216 (2015) 574–578.
- [7] G. Hripcsak, P.B. Ryan, J.D. Duke, et al., Characterizing treatment pathways at scale using the OHDSI network, *Proc. Natl. Acad. Sci.* 113 (27) (2016) 7329–7336.
- [8] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, N.D. Lawrence, Dataset shift in machine learning, The MIT Press, 2009.
- [9] J. Chen, L. Sun, C. Guo, Y. Xie, A fusion framework to extract typical treatment patterns from electronic medical records, *Artif. Intell. Med.* 103 (2020), 101782.
- [10] F. Johansson, U. Shalit, D. Sontag, Learning representations for counterfactual inference, in: *International Conference on Machine Learning*, 2016, pp. 3020–3029.
- [11] L. Yao, S. Li, Y. Li, M. Huai, et al., Representation learning for treatment effect estimation from observational data, *Adv. Neural Inf. Process. Syst.* (2018) 2633–2643.
- [12] P. Chen, W. Dong, X. Lu, et al., Deep representation learning for individualized treatment effect estimation using electronic health records, *J. Biomed. Inform.* 100 (2019), 103303.
- [13] E. Choi, A. Schuetz, W. Stewart, J. Sun, Using recurrent neural network models for early detection of heart failure onset, *J. Amer. Med. Informat. Assoc.* 24 (2) (2017) 361–370.
- [14] J. Chu, W. Dong, Z. Huang, Endpoint prediction of heart failure using electronic health records, *J. Biomed. Inform.* 109 (2020), 103518.
- [15] H. Duan, Z. Sun, W. Dong, K. He, Z. Huang, On clinical event prediction in patient treatment trajectory using longitudinal electronic health records, *IEEE J. Biomed. Health. Inf.* 24 (7) (2020) 2053–2063.
- [16] Z. Huang, W. Dong, L. Ji, H. Duan, Outcome prediction in clinical treatment processes, *J. Med. Syst.* 40 (1) (2016) 8.
- [17] D. Yoon, E.K. Ahn, M.Y. Park MY, et al., Conversion and data quality assessment of electronic health record data at a Korean tertiary teaching hospital to a common data model for distributed network research, *Healthc. Inform. Res.* 22(1) (2016) 54–58.
- [18] E. Voss, R. Makadia, A. Matcho, et al., Feasibility and utility of applications of the common data model to multiple, disparate observational health databases, *J. Am. Med. Inform. Assoc.* 22 (3) (2015) 553–564.
- [19] A. Ostroplets, C. Reich, P. Ryan, et al., Adapting electronic health records-derived phenotypes to claims data: Lessons learned in using limited clinical data for phenotyping, *J. Biomed. Inform.* 102 (2020), 103363.
- [20] C. Weng, N.H. Shah, G. Hripcsak, Deep phenotyping: Embracing complexity and temporality—Towards scalability, portability, and interoperability, *J. Biomed. Inform.* 105 (2020), 103433.
- [21] X. Lv, Y. Guan, B. Deng, Transfer learning based clinical concept extraction on data from multiple sources, *J. Biomed. Inform.* 52 (2014) 55–64.
- [22] Y. Gu, Z. Ge, C.P. Bonnington, J. Zhou, Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification, *IEEE J. Biomed. Health. Inf.* 24 (5) (2019) 1379–1393.
- [23] P.C. Austin, An introduction to propensity score methods for reducing the effects of confounding in observational studies, *Multivar. Behav. Res.* 46 (3) (2011) 399–424.
- [24] R.H. Dehejia, S. Wahba, Propensity score-matching methods for nonexperimental causal studies, *Rev. Econ. Stat* 84 (1) (2002) 151–161.
- [25] P.R. Rosenbaum, D.B. Rubin, The central role of the propensity score in observational studies for causal effects, *Biometrika* 70 (1) (1983) 41–55.
- [26] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, *IEEE TPAMI* 35 (8) (2013) 1798–1828.
- [27] X. Xu, X. Zhou, F. Shen, et al., Fusion by synthesizing: A multi-view deep neural network for zero-shot recognition, *Signal Process.* 164 (2019) 354–367.
- [28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., Generative adversarial nets, *Proc. Adv. Neural Inf. Process. Syst.* (2014) 2672–2680.
- [29] Z. Huang, W. Dong, Adversarial MACE prediction after acute coronary syndrome using electronic health records, *IEEE J. Biomed. Health. Inf.* 23 (5) (2019) 2117–2126.
- [30] J. Yoon, J. Jordon, M. van der Schaar, GANITE: Estimation of individualized treatment effects using generative adversarial nets, in: *International Conference on Learning Representations*, 2018.
- [31] H. Boshnak, S. AbdelGaber, A. Abdou, E. Yehia, Ontology-based knowledge modelling for clinical data representation in electronic health records, *Int. J. Comput. Sci. Inf. Security* 16 (2019) 68–86.
- [32] D. Chen, S. Liu, P. Kingsbury, et al., Deep learning and alternative learning strategies for retrospective real-world clinical data, *npj Digital Med.* 2 (1) (2019) 1–5.
- [33] M. Valgimigli, F. Saia, P. Guastaroba, et al., Transradial versus transfemoral intervention for acute myocardial infarction: a propensity score-adjusted and-matched analysis from the REAL (Registro regionale Angioplastiche dell'Emilia-Romagna) multicenter registry, *JACC: Cardiovasc. Intervent.* 5 (1) (2012) 23–35.
- [34] R. Pirracchio, M. Resche-Rigon, S. Chevret, Evaluation of the Propensity score methods for estimating marginal odds ratios in case of small sample size, *BMC Med. Res. Methodol.* 12 (1) (2012) 70.
- [35] G.N. Okoli, R.D. Sanders, P. Myles, Demystifying propensity scores, *Br. J. Anaesth.* 112 (1) (2014) 13–15.
- [36] D.L. Streiner, G.R. Norman, The pros and cons of propensity scores, *Chest* 142 (6) (2012) 1380–1382.
- [37] Z. Zhang, C. Yan, D.A. Mesa, J. Sun, B.A. Malin, Ensuring electronic medical record simulation through better training, modeling, and evaluation, *J. Am. Med. Inform. Assoc.* 27 (1) (2020) 99–108.
- [38] F. Wang, J. Sun, S. Ebadollahi, Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment, *Stat. Anal. Data Min.* 5 (1) (2012) 54–69.
- [39] J. Sun, F. Wang, J. Hu, S. Ebadollahi, Supervised patient similarity measure of heterogeneous patient records, *ACM SIGKDD Explor. Newslett.* 14 (1) (2012) 16–24.
- [40] W. Zellinger, T. Grubinger, E. Luehner, T. Natschlager, S. Saminger-Platz, Central moment discrepancy (CMD) for domain-invariant representation learning, *arXiv preprint arXiv:1702.08811*, 2017.
- [41] S. Pokharell, G. Zuccon, X. Li, C.P. Utomo, Y. Li, Temporal tree representation for similarity computation between medical patients, *Artif. Intell. Med.* (2020), 101900.
- [42] S. Ma, K. Yang, N. Wang, et al., Disease phenotype synonymous prediction through network representation learning from PubMed database, *Artif. Intell. Med.* 102 (2020), 101745.
- [43] A.J. Averitt, N. Vanitchanan, R. Ranganath, A.J. Perotte, The counterfactual χ -GAN: finding comparable cohorts in observational health data, *J. Biomed. Inform.* 109 (2020), 103515.
- [44] I. Bica, A.M. Alaa, J. Jordon, M. van der Schaar, Estimating counterfactual treatment outcomes over time through adversarially balanced representations, in: *International Conference on Learning Representations (ICLR)*, 2020.
- [45] L. Li, P. Wang, J. Yan, et al., Real-world data medical knowledge graph: construction and applications, *Artif. Intell. Med.* 103 (2020), 101817.
- [46] S. Hong, C. Xiao, T. Ma, H. Li, J. Sun, MINA: Multilevel knowledge-guided attention for modeling electrocardiography signals, in: *International Joint Conferences on Artificial Intelligence*, 2019, pp. 5888–5894.
- [47] G. Bakal, P. Talari, E.V. Kakani, R. Kavuluru, Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations, *J. Biomed. Inform.* 82 (2018) 189–199.
- [48] E. Choi, M.T. Bahadori, L. Song, et al., GRAM: graph-based attention model for healthcare representation learning, in: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*, 2017, pp. 787–795.
- [49] S. Bickel, M. Bruckner, T. Scheffer, Discriminative learning under covariate shift, *J. Mach. Learn. Res.* 10 (2019) 2137–2155.
- [50] B. Yang, W. Yih, X. He, J. Gao, L. Deng, Embedding Entities and Relations for Learning and Inference in Knowledge Bases, *arXiv:1412.6575*, 2014.
- [51] A. Bordes, J. Weston, R. Collobert, Y. Bengio, Learning structured embeddings of knowledge bases, in: *AAAI*, 2011.
- [52] A. Bordes, X. Glorot, J. Weston, Y. Bengio, A semantic matching energy function for learning with multi-relational data, *Mach. Learn.* 94 (2) (2013) 233–259.
- [53] A. Bordes, N. Usunier, A. Garcia-Duran, et al., Translating embeddings for modeling multi-relational data, *Adv. Neural Inf. Process. Syst.* 26 (2013) 2787–2795.
- [54] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: *3rd International Conference on Learning Representations*, *arXiv preprint arXiv:1412.6980*, 2014.
- [55] J. Wang, X. Shen, W. Pan, On transductive support vector machines, *Contemp. Math.* 443 (2007) 7–20.
- [56] T. Joachims, Transductive inference for text classification using support vector machines, *ICML* 99 (1999) 200–209.
- [57] D.P. Kingma, S. Mohamed, D.J. Rezende, M. Welling, Semi-supervised learning with deep generative models, *Adv. Neural Inf. Process. Syst.* 27 (2014) 3581–3589.
- [58] I. Anand, J.J.V. McMurray, J. Whitmore, et al., Anemia and its relationship to clinical outcome in heart failure, *Circulation* 110 (2) (2004) 149–154.