




A decision support system in precision medicine: contrastive multimodal learning for patient stratification

Qing Yin¹ · Linda Zhong² · Yunya Song³ · Liang Bai⁴ · Zhihua Wang⁵ · Chen Li⁶ · Yida Xu⁷ · Xian Yang¹ 

Received: 29 November 2022 / Accepted: 8 August 2023 / Published online: 29 August 2023
© The Author(s) 2023

Abstract

Precision medicine aims to provide personalized healthcare for patients by stratifying them into subgroups based on their health conditions, enabling the development of tailored medical management. Various decision support systems (DSSs) are increasingly developed in this field, where the performance is limited to their capability of handling big amounts of heterogeneous and high-dimensional electronic health records (EHRs). In this paper, we focus on developing a deep learning model for patient stratification that can identify and explain patient subgroups from multimodal EHRs. The primary challenge is to effectively align and unify heterogeneous information from various modalities, which includes both unstructured and structured data. Here, we develop a **Contrastive Multimodal learning model for EHR (ConMEHR)** based on topic modelling. In ConMEHR, modality-level and topic-level contrastive learning (CL) mechanisms are adopted to obtain a unified representation space and diversify patient subgroups, respectively. The performance of ConMEHR will be evaluated on two real-world EHR datasets and the results show that our model outperforms other baseline methods.

Keywords Modelling unstructured and structured patient data · Application of EHRs in precision medicine · Deep learning model for patient stratification · Multimodal contrastive learning

1 Introduction

Electronic health records (EHRs) are becoming increasingly valuable in precision medicine as they provide diverse and comprehensive information on patients' health conditions, which can facilitate clinical decision-making using decision support systems (DSSs). EHRs comprise a wide range of information from various sources, such as patient demographics, diagnoses, laboratory test results, prescribed medications, clinical notes, and medical images (Johnson et al., 2016). The explosive growth of EHRs has created an urgent need for innovative methods in DSSs that can efficiently utilize large amounts of high-dimensional EHR data to stratify

✉ Xian Yang
xian.yang@manchester.ac.uk

Extended author information available on the last page of the article

patients. Patient stratification plays a crucial role in advancing precision medicine and personalized treatment. By stratifying patients into subgroups based on their health conditions, healthcare providers can better understand each individual's specific medical needs, allowing for the development of customized healthcare interventions that minimize risk. Furthermore, patient stratification can help identify patients who are likely to benefit from particular treatments, enabling healthcare providers to make informed decisions and deliver personalized care that is optimized for each patient.

Patient stratification refers to the identification of subgroups of patients with similar health conditions. It is important to note that, for the sake of explainability, patient stratification also explains each subgroup using terms such as disease severity, phenotype, and diagnosis that are commonly shared among patients within the same group. This task is illustrated in Fig. 1, where the input EHR is represented as a patient-word matrix. In this matrix, the 0/1 value in the i -th row and j -th column indicates the absence/presence of the j -th word (e.g., clinical concept) in patient i . By applying computational methods, the input matrix is transformed into two matrices: the patient-subgroup matrix and the subgroup-word matrix. Each row in the patient-subgroup matrix indicates the probability of a patient being assigned to different subgroups, while each row in the subgroup-word matrix explains each subgroup by a distribution of words. Using these two matrices, we can detect the potential patients and words in each subgroup, respectively.

Conventional approaches for patient stratification often use topic modelling to extract explainable topics as patient subgroups. Latent Dirichlet Allocation (LDA) (Blei, 2012) is a widely used probabilistic modelling technique for discovering latent topics from EHR data. LDA assumes that each EHR is a mixture of some topics. Recent approaches to improving the performance of LDA have combined topic modelling with neural networks. For example, the neural topic model (NTM) (Zhao et al., 2021) uses neural variational inference to create parameterized topic distributions during training. One of the challenges of NTM is to obtain good topics when there are large narratives and big vocabulary sizes. The embedded topic model (ETM) (Dieng et al., 2020), a recent approach to NTM, overcomes this issue by integrating neural topical modelling with word embeddings (Mikolov et al., 2013). This generative model assumes that each document (e.g., EHR) is produced from a blend of topics, and each observed word (e.g., clinical concept) in the document is drawn from some topics. All topics and observed words are embedded with numeric vectors, and similarities between these vectors can indicate the similarities among words and topics.

In this paper, our focus is on developing a patient stratification model that leverages EHRs with multiple modalities. However, one challenge of applying the aforementioned methods to model multimodal EHRs is how to align and unify heterogeneous information from different modalities for learning consensus topics. Common multimodal data integration

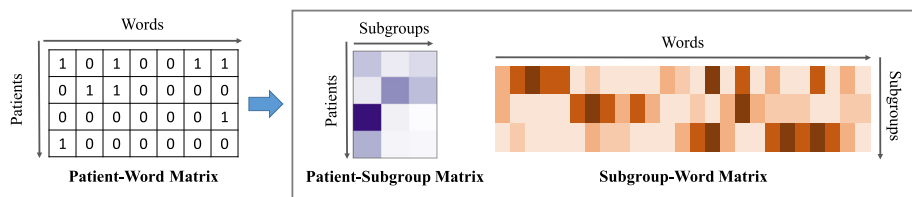


Fig. 1 The illustration of patient stratification. The patient-word matrix is the representation of the raw input EHR. With the application of patient stratification models, the input data can be decomposed into two sub-matrices: the patient-subgroup matrix and the subgroup-word matrix (darker colours indicate larger values). (Color figure online)

methods include early, hybrid, and late fusions, where the work in Xu et al. (2021) adopted the neural architecture search to find an optimal data fusion strategy. Moreover, there are increasing efforts being made nowadays to ensure different modalities are encoded into a unified semantic space. Among them, multi-modality contrastive learning (CL) (Grill et al., 2020; Oord et al., 2018; Caron et al., 2020; Li et al., 2021) has shown great potential. For example, Coca (Yu et al., 2022) utilized a dual encoder model like CLIP (Radford et al., 2021) to fuse image and text data, while FLAVA (Singh et al., 2022) aligned multimodal data at a fine-grain level via cross-attention. All these methods adopted the idea of CL by augmenting data with positive and negative sample pairs. However, most of them are not specifically designed to handle multimodal data from EHRs.

Here, we develop a **Contrastive Multimodal learning model for EHR** (ConMEHR) for patient stratification. Our study is centred on the two most frequently encountered types of modalities in EHRs: unstructured texts and structured medical terms. Unstructured texts refer to free-form narrative data, such as medical notes and reports, while structured medical terms are structured data elements containing standard medical terms like the International Classification of Diseases (ICD) codes (Organization et al., 1978) and medication names. For illustration purposes, we provide information from four representative modalities of these two types found in EHRs: Note, Disease, Symptom, and Medication. Note is an unstructured text modality, while Disease, Symptom, and Medication are the structured medical term modalities. ConMEHR addresses the challenge of unifying heterogeneous information from diverse modalities by utilizing modality-level CL. This approach enables the learning of a unified latent representation space for various modalities, which can then be integrated to obtain patient subgroups. As a topic modelling based approach, the performance of ConMEHR would also be influenced by the ability to learn diverse topics (Ding et al., 2015). Previous works examined the topic diversity by identifying unique words associated with each topic (Benson et al., 2014; Arora et al., 2013) while ignoring the semantic-level separation of topics. Hence, to address this issue, we consider imposing a topic-level CL module to diversify topics by separating their latent representations.

To demonstrate the applicability of ConMEHR, we use two EHR datasets: MIMIC-III, a publicly accessible EHR dataset in English, and a Chinese EHR dataset collected from Chinese medical clinics. To encode EHR characters in English and Chinese, we utilized BioBERT (Lee et al., 2020) and Chinese-BERT (Cui et al., 2021), respectively. We focus on four representative modalities of EHRs, namely, Note, Disease, Symptom, and Medication, to illustrate our approach. Considering the difficulty of modelling data with high complexity, we incorporate a regularization module as a training trick in our model. Our regularization module is based on the one proposed in RDrop (Wu et al., 2021). This module is trained by minimizing the bidirectional Kullback–Leibler (KL) divergence between the output distributions of two sub-models after dropout. It reduces the freedom of the model parameters and hence improves the model's adaptability to high-complexity multiple modalities data. Our contributions in this paper can be summarized as follows:

- We propose ConMEHR, a deep learning model for patient stratification that utilizes both unstructured and structured modalities of EHRs to identify patient subgroups.
- To effectively integrate diverse information from multiple modalities, we propose using modality-level and topic-level contrastive learning mechanisms to obtain a unified latent representation space for the various modalities, while also diversifying patient subgroups.
- We evaluate ConMEHR on two real-world EHR datasets, using both quantitative and qualitative methods. Case studies, such as patient subgroup interpretation and visualization, demonstrate the significant power of our model in DSS.

2 Related work

2.1 Patient stratification using EHRs

Patient stratification using EHRs is a crucial task in precision medicine that provides decision support. This task involves grouping patients into different categories and explaining each subgroup using shared terms among patients in the same group. This approach allows clinicians and researchers to gain insights into the underlying factors contributing to a specific patient's condition and provide accurate prognoses. This is in contrast to classification models such as Bagging CART and XGBoost, which only focus on predicting patients to predefined classes. There are primarily two branches of methods used for patient stratification: tensor factorization and topic modelling.

Tensor factorization (Kim et al., 2017; Wang et al., 2015) is an efficient approach for patient stratification, where massive EHRs are converted into meaningful concepts. For example, PARAFAC2 (Harshman, 1972) automated the stratification for patients who need intensive medical care. Additionally, in Henderson et al. (2017), a diversified and sparse nonnegative tensor factorization method was developed to derive patient subgroups from EHRs. Another work in Ho et al. (2014) stratified patients by decomposing the tensor representation of EHR using the CP-APR algorithm (Chi & Kolda, 2012).

Topic models are commonly used to deal with unstructured texts and explain patient groups by learning topics so that patients with similar observations can be grouped together based on shared topics. LDA (Blei, 2012) is a classical topic modelling approach that uses bag-of-words (BoW) inputs to find patient subgroups from documents. There have been numerous efforts to enhance its performance. For example, methods introduced in (Shi et al., 2017; Zhao et al., 2017) modify the prior distributions over topics in LDA. In Xun et al. (2017), a topic model is incorporated with word embedding to first convert the discrete texts into continuous representations. The work in Bunk and Krestel (2018) involved replacing topic-drawn words with Gaussian-distributed embeddings in a random manner. The emergence of neural networks has led to a surge in the number of approaches that combine probabilistic topic models with deep neural networks (Srivastava & Sutton, 2017; Cong et al., 2017; Zhang et al., yyyy). The majority of these approaches employ variational auto-encoder and amortized inference to reduce input data dimensionality (Rezende et al., 2014; Dieng et al., 2020). Among them, ETM (Dieng et al., 2020) and its extensions (Zou et al., 2022; Wang et al., 2022) are neural topic models that use word embeddings and also learn topic embeddings. In our study, the technical framework is based on the ETM approach, which leverages the explanatory power of topic modelling to explain subgroups while also incorporating semantic features through the use of embedding representations.

2.2 Contrastive representation learning

CL is a type of self-supervised learning that has gained increasing attention due to its tremendous impact on representation learning (Mikolov et al., 2013; Chopra et al., 2005; Logeswaran & Lee, 2018; Weinberger & Saul, 2009; Chechik et al., 2010; Hoffer & Ailon, 2015; Oh Song et al., 2016; Mikolov et al., 2013; Henaff, 2020; Hjelm et al., 2018; Wu et al., 2018; He et al., 2020; Zhang et al., 2022). CL works based on the InfoMax principle (Linsker, 1988) by maximizing the mutual information between two augmented views of a sample (e.g., an image with different rotations and shifts) (Tian et al., 2020; Bachman et al., 2019). In CL, an anchor and a positive sample should be drawn closer together, while an anchor and a negative

sample should be drawn farther apart (Chen et al., 2020; He et al., 2020; Pan et al., 2021; Gao et al., 2021).

One of the research areas of interest for CL is the design of positive pairs (Arora et al., 2019; Gao et al., 2021). For single-modality CL approaches, a typical practice is to optimize by an auxiliary set gained through data augmentation (Wu et al., 2018; Ho & Nvasconcelos, 2020; He et al., 2020; Chen et al., 2020). These methods disperse different instances apart while implicitly bringing similar instances together. In multi-modality CL approaches, for each modality of a certain sample, other modalities from the same sample will be regarded as positives, while any modalities from different samples would be negatives (Grill et al., 2020; Oord et al., 2018; Caron et al., 2020; Li et al., 2021; You et al., 2021; Zhang et al., 2022). These methods guarantee that the features from different modalities of the same sample map to proximate points in the latent representation space. Among them, MCSE (Zhang et al., 2022) is the state-of-the-art multimodal contrastive model used for learning sentence embeddings. It employs a multimodal contrastive objective that aligns data from different modalities in an embedding space. This generic multimodal objective can be integrated into various sentence embedding techniques, potentially improving their effectiveness.

3 Methodology

3.1 Patient stratification system

As shown in Fig. 2, the proposed patient stratification system consists of three key steps, which are data preprocessing, the ConMEHR model and applications in precision medicine.

The data preprocessing is responsible for receiving heterogeneous data from EHR data warehouses and preparing it for subsequent analyses. In the patient stratification system presented in this paper, the module takes information from four representative modalities of EHRs: Note (*N*), Disease (*D*), Symptom (*S*), and Medication (*M*). However, it can be naturally extended in the future to handle other modalities. Among these modalities, Disease, Symptom, and Medication are structured medical term modalities that consist of structured data. In contrast, Note contains unstructured texts that are more likely to have noisy informa-

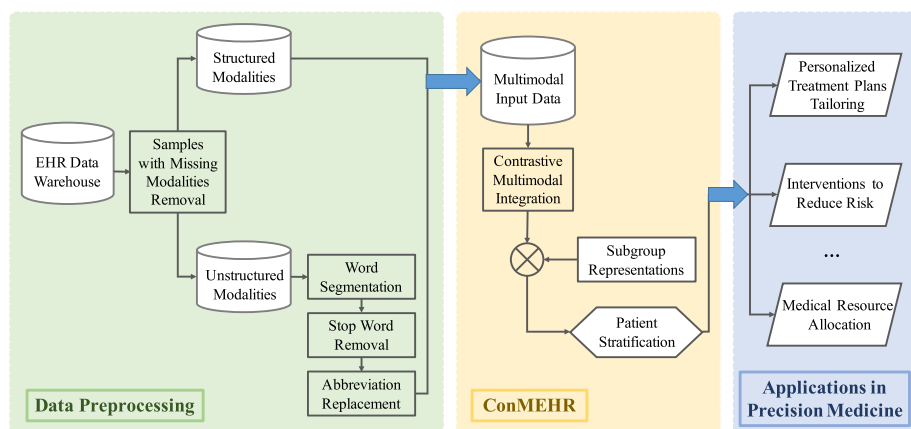


Fig. 2 The pipeline of our patient stratification system

tion resulting from synonyms, abbreviations, and informal descriptions. Therefore, a range of natural language processing (NLP) techniques, such as word segmentation, stop word removal, and abbreviation replacement methods will be adopted to preprocess the unstructured texts. Word segmentation involves breaking down lengthy texts, such as paragraphs and articles, into word units to facilitate easier analysis and processing. For English text, we usually split it according to space, while for Chinese text with different written rules, we use the integrated toolkit *jieba*¹ to split sentences. Stop words, which are frequently used words such as ‘the’, ‘a’, ‘an’, or ‘in’ are ignored by search engines. Since we do not intend to utilize these words in our modelling and analysis processes, we can easily remove them by creating a list of potential stop words. We utilize the Natural Language Toolkit (*NLTK*)² in Python, which includes stopping word lists in 16 different languages. For abbreviation replacement, there are mainly two steps: abbreviation detection and replacement. The *AbbreviationDetector* component in *Spacy*,³ a free open-source library for NLP in Python, can support abbreviation detection. We apply *AbbreviationDetector* to identify abbreviations and replace them with their full names.

The ConMEHR model is used to stratify patients based on preprocessed multimodal EHR data. This approach extends the idea of topic modelling, which assumes that patients belong to latent topics or subgroups, and the patient-to-topic distribution can determine the probabilities of assigning patients to different subgroups. To achieve this, we use the ETM model, which employs deep learning in representation learning to generate the representations of learned latent topics, ensuring explainable results. ETM is used as our backbone model and improved to integrate the heterogeneous information from multiple modalities of EHRs via CL. To encode each modality, neural networks first project raw data into a latent representation space. It is important to note that multiple modalities from the same EHR are potentially correlated and can provide complementary information to each other. Therefore, to ensure that their latent representations share some common characteristics, we propose modality-level CL. This technique is widely used to obtain robust representations in the manner of self-supervised learning, and we will explain how we adapt it to our problem in the following section. In addition to modality-level CL, we also design topic-level CL to diversify topics or subgroups and learn separable topic representations.

The results of the ConMEHR model are promising for various applications in precision medicine, including personalized treatment plans, targeted interventions to reduce individual risk, and efficient allocation of medical resources. In the following subsection, we will provide detailed information about the ConMEHR model.

3.2 Our ConMEHR model

As shown in Fig. 3, ConMEHR consists of two modelling processes: the inference process and the generative process. We will first describe the inference process to get the patient subgroup distributions from the multimodal EHR data via modality-level CL. Then, we introduce the generative process which adopts topic-level CL to get distinguished subgroup representations via BoW reconstruction. Finally, the optimization objective used in the model training will be defined. The main notations used in this section are summarized in Table 1.

¹ <https://github.com/fxsjy/jieba>

² <https://www.nltk.org/>

³ <https://spacy.io/>

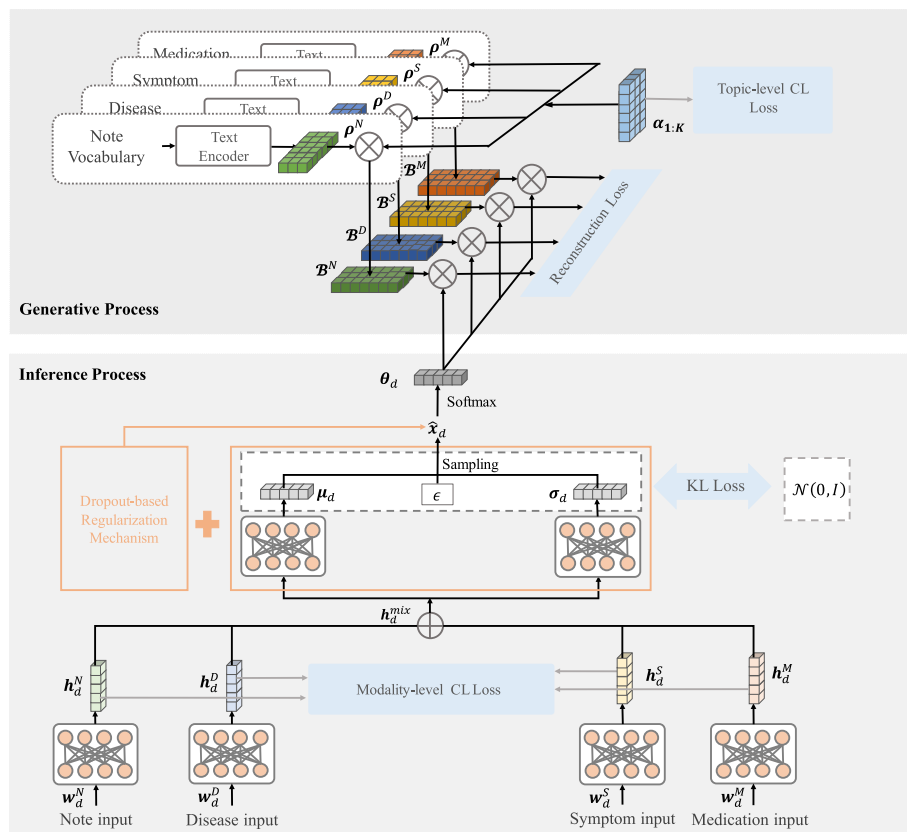


Fig. 3 The structure of the ConMEHR model. For the d -th EHR, the inference process encodes its four modalities to get the topic distribution θ_d . The four modalities are Note, Disease, Symptom, and Medication, labelled as N , D , S , and M , respectively. The text encoder in the generative process is BioBERT or ChineseBERT which depends on whether the input data is Chinese or English text. The modality-level CL is used to align and unify those four modalities in the same semantic space. The dropout-based regularization module is used here to improve the model's representation ability via dropout technology. The generative process uses the topic distribution θ_d and embeddings of vocabularies to learn subgroup representations and reconstruct the BoW representation of inputs. The topic-level CL is used to get separable topic representations

3.2.1 Inference process for patient subgrouping through contrastive multimodal learning

For each EHR d , the inference process encodes its multimodal information to approximate the posterior distribution of the latent representation x_d . The posterior distribution is denoted as $q(x_d | w_d^*)$, where $w_d^* = \{w_d^N, w_d^D, w_d^S, w_d^M\}$ contains the normalized BoW representations of four modalities. For each modality, a separated MultiLayer Perceptron (MLP) module is employed to map the normalized BoW inputs into the latent space. Let h_d^N, h_d^D, h_d^S , and h_d^M denote the latent representations of four modalities respectively. One challenge of integrating these latent representations lies in aligning and unifying them in the same semantic space.

Inspired by the work in Li et al. (2021), we introduce the modality-level CL mechanism to facilitate multimodal data integration. CL is a self-supervised learning approach for learning

Table 1 Symbols and descriptions

Symbols	Descriptions
B	The batch size used in the training process
d	The index of EHR, $d \in \{1, 2, \dots, B\}$
K	The total number of topics
k	The index of topic
L	The dimension of the embedding space
N, D, S, M	Notations of four modalities: N for Note, D for Diagnosis, S for Symptom, and M for Medication
t	The indicator of modality, $t \in \{N, D, S, M\}$
N_d^t	The total number of words in modality t from the d -th EHR
$\mathbf{w}_d^t \in \mathbb{R}^{N_d^t}$	The normalized BoW representation of modality t from the d -th EHR
\mathbf{w}_d^*	$\mathbf{w}_d^* = \{\mathbf{w}_d^N, \mathbf{w}_d^D, \mathbf{w}_d^S, \mathbf{w}_d^M\}$
$w_{d,n}^t$	The n -th word in modality t from the d -th EHR
V_t	The vocabulary of modality t across all samples
$ V_t $	The vocabulary size of modality t
$z_{d,n}^t$	The topic assignment of $w_{d,n}^t$
$\mathbf{h}_d^t \in \mathbb{R}^L$	The latent representation of modality t from the d -th EHR
$\mathbf{h}_d^{\text{mix}} \in \mathbb{R}^{4L}$	The integrated representation for the d -th EHR
$\alpha_k \in \mathbb{R}^L$	The embedding vector of topic k
$\alpha_{1:K} \in \mathbb{R}^{L \times K}$	The embedding matrix of all K topics
$\beta_k^t \in \mathbb{R}^{ V_t }$	The modality t 's word probability distribution for topic k
$\mathcal{B}^t \in \mathbb{R}^{ V_t \times K}$	The modality t 's word probability distributions for all topics, $\mathcal{B}^t = [\beta_1^t, \dots, \beta_k^t, \dots, \beta_K^t]$
$\theta_d \in \mathbb{R}^K$	The topic distribution for the d -th EHR
$\mathbf{x}_d \in \mathbb{R}^K$	The latent representation of the d -th EHR, from which θ_d will be generated
$\rho^t \in \mathbb{R}^{L \times V_t }$	The embedding matrix of modality t 's vocabulary
$\mu_d \in \mathbb{R}^K$	The mean vector of \mathbf{x}_d
$\sigma_d \in \mathbb{R}^K$	The standard deviation vector of \mathbf{x}_d

robust representations by encouraging augmented samples of the same input to have relatively similar representations to other augmented samples. It is based on the concept of positive and negative samples in relation to an anchor point. This entails pulling the anchor and positive sample closer in the embedding space while simultaneously pushing the anchor away from multiple negative samples. The modality-level CL mechanism we proposed is to encourage representations from different modalities of the same sample to become more similar than those from different samples. To adopt the modality-level CL mechanism, we construct positive and negative pairs for the training. The positive pair $(\mathbf{h}_i^t, \mathbf{h}_i^{t'})$ is constructed by selecting an anchor point \mathbf{h}_i^t from one modality t and pairing it with $\mathbf{h}_i^{t'}$, which is derived from the same EHR i but belongs to a different modality. The negative pairs can be classified into two distinct types. The first type consists of $(\mathbf{h}_i^t, \mathbf{h}_j^t)$ pairs, where $j \neq i$ represents a different EHR index within the current batch, but both vectors belong to the same modality

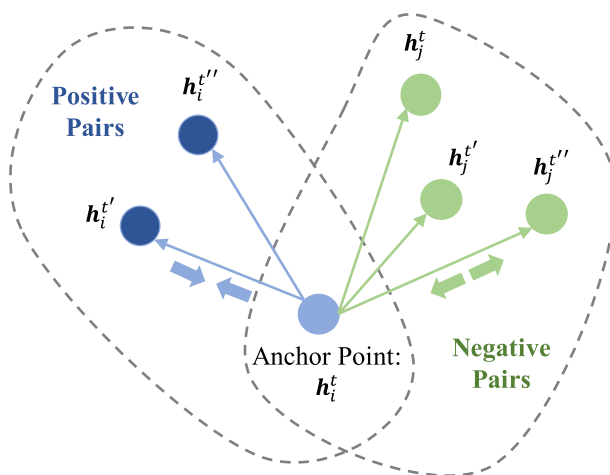


Fig. 4 The illustration of modality-level CL. In the embedding space, for an anchor point h_i^t , the modality-level CL aims to pull positive samples like $h_i^{t'}$ and $h_i^{t''}$ closer and push apart negative ones like h_j^t , $h_j^{t'}$ and $h_j^{t''}$. Here, $h_i^{t'}$ and $h_i^{t''}$ are from the EHR i same as the anchor h_i^t but of different modalities, while h_j^t , $h_j^{t'}$ and $h_j^{t''}$ are from a different EHR j

t . The second type of negative pair comprises $(h_i^t, h_j^{t'})$, where $h_j^{t'}$ is distinct not only in the EHR index but also in the modality index from the anchor point h_i^t . Specifically, for a batch of EHRs, the modality-level contrastive learning loss is as shown in Fig. 4 and defined as follows:

$$\mathcal{L}_{mcl} = - \sum_{t=1}^4 \sum_{t' \geq t}^4 \sum_{i=1}^B \log \frac{\exp(\text{sim}(h_i^t, h_i^{t'})/\tau)}{\sum_{j=1}^B \mathbb{1}_{j \neq i} \cdot \exp(\text{sim}(h_i^t, h_j^{t'})/\tau)}, \quad (1)$$

where t and t' are the indices of modalities, B is the batch size, and $\mathbb{1}_{j \neq i}$ is an indicator function which equals to 1 when $j \neq i$ and equals to 0 when $j = i$. τ denotes the temperature parameter which is used to regulate the severity of penalties imposed on the hard negative samples. $\text{sim}(\cdot)$ measures the cosine similarity between two vectors (Chen et al., 2020). h_i^t is the latent representations of modality t from the i -th EHR. The numerator of Eq.(1) represents the similarity of two modalities of the same sample, while the denominator represents the similarity of two modalities from different samples in the batch.

With the constraints imposed by the modality-level CL, four latent representations are concatenated into one vector via:

$$h_d^{mix} = h_d^N \oplus h_d^D \oplus h_d^S \oplus h_d^M, \quad (2)$$

where \oplus is the concatenation operator. h_d^{mix} can be regarded as an integrated representation of the multimodal EHR. Then, we approximate the posterior distribution of x_d using h_d^{mix} . Following the framework of neural variational inference approaches (Miao et al., 2016; Kingma & Welling, 2013; Rezende et al., 2014), the posterior $q(x_d | w_d^*)$ is assumed to be Gaussian and its mean vector μ_d and standard deviation vector σ_d are approximated via two separated MLP modules respectively. A sample of x_d can be obtained from $q(x_d | w_d^*)$ as follows:

$$\hat{x}_d = \mu_d + \epsilon \cdot \sigma_d^2 \quad (3)$$

where $\epsilon \in \mathcal{N}(0, \mathbf{I})$. With $\hat{\mathbf{x}}_d$, the per EHR topic distribution $\boldsymbol{\theta}_d$ can be generated as:

$$\boldsymbol{\theta}_d = \text{softmax}(\hat{\mathbf{x}}_d), \quad (4)$$

where $\text{softmax}(\cdot)$ is the Softmax activation function (Nwankpa et al., 2021) to transform the Gaussian sample $\hat{\mathbf{x}}_d$ into the topic distribution $\boldsymbol{\theta}_d$.

3.2.2 Generative process for subgroup representation learning via BoW reconstruction

Similar to our backbone model ETM, the generative process in the ConMEHR model utilizes the embeddings of vocabularies and also learns the embedding matrix of topics. Each word is assigned with an L -dimensional embedding vector by pretrained language models (PLMs). PLMs are extensive neural networks that find application in a broad range of NLP tasks. These models operate using a pretrain-finetune approach, where they undergo pretraining on a vast corpus of text, following which they are refined for a downstream task. PLMs are considered to be effective language encoders as they offer fundamental language comprehension abilities that can be leveraged across various downstream applications. Elazar et al. (2021). Bidirectional Encoder Representations from Transformers (BERT) (Kenton & Toutanova, 2019) is one classical PLM that utilizes Transformer (Vaswani et al., 2017), a novel neural network architecture based on a self-attention mechanism, for language understanding. There are several different BERT models available. For example, Chinese-BERT (Cui et al., 2021) is a language model pretrained on a large Chinese text corpus for encoding Chinese texts. BioBERT (Lee et al., 2020) is a pretrained biomedical language model which adopts the architecture of BERT (Devlin et al., 2019) and is fine-tuned on PubMed⁴ abstracts and PMC articles. SentenceBERT (Reimers & Gurevych, 2019) is a language model which has fine-tuned BERT (Devlin et al., 2019) for better measuring sentence similarities. In our paper, we focus on using Chinese-BERT⁵ and BioBERT⁶ to encode terms in the Chinese EHR dataset and the English EHR dataset respectively.

Suppose the latent representation of each topic is represented as an embedding vector $\boldsymbol{\alpha}_k \in \mathbb{R}^L$ for $k \in \{1, \dots, K\}$, where K is the total number of topics. In our model, $\boldsymbol{\alpha}_k$ will be randomly initialized and updated via end-to-end training. To learn separable topic representations, we follow the work in Zhang et al. (2021) and introduce topic-level CL. In our topic-level CL module, for each topic k we regard its embedding vector $\boldsymbol{\alpha}_k$ as the anchor and the embeddings from other topics as negatives. Specifically, positive pairs are formed by using identical embedding vectors, where both vectors in the positive pair $(\boldsymbol{\alpha}_k, \boldsymbol{\alpha}_k)$ correspond to the same topic k . On the other hand, negative pairs are formed by using non-identical embedding vectors and consist of pairs $(\boldsymbol{\alpha}_k, \boldsymbol{\alpha}_{k'})$ where k and k' represent different topics. The topic-level contrastive learning loss is defined as follows:

$$\mathcal{L}_{tcl} = - \sum_{k=1}^K \log \frac{\exp(\text{sim}(\boldsymbol{\alpha}_k, \boldsymbol{\alpha}_k)/\tau)}{\sum_{k'=1}^K \mathbb{1}_{k' \neq k} \cdot \exp(\text{sim}(\boldsymbol{\alpha}_k, \boldsymbol{\alpha}_{k'})/\tau)}, \quad (5)$$

where the numerator of Eq. (5) is a constant value equal to $\exp(1/\tau)$, while the denominator represents the similarity of the samples in negative pairs. Therefore, Eq. (5) can be rewritten as $\mathcal{L}_{tcl} = -(1/\tau) + 2 \sum_{k=1}^K \log \sum_{k' > k} \exp(\text{sim}(\boldsymbol{\alpha}_k, \boldsymbol{\alpha}_{k'})/\tau)$, implying that the training goal is to distinguish representations from different topics.

⁴ <https://www.ncbi.nlm.nih.gov/pmc/>

⁵ <https://github.com/ymcui/Chinese-BERT-wwm>

⁶ <https://github.com/dmis-lab/biobert>

In the generative process, the BoW representations of EHRs are reconstructed using all modalities' word probability distributions for all topics obtained from the embeddings of vocabularies and topics. The detailed steps are described as follows. For each EHR d , using the topic distribution θ_d from Eq. (4) in the inference process, the topic assignment of the n -th word in modality t is drawn from:

$$z_{d,n} \sim \text{Cat}(\theta_d), \quad (6)$$

where $\text{Cat}(\cdot)$ refers to the categorical distribution and $z_{d,n} \in \{1, 2, \dots, K\}$. The word probability distribution for the assigned topic $z_{d,n}$ can be obtained from:

$$\beta_{z_{d,n}}^t = \text{softmax}((\rho^t)^T \alpha_{z_{d,n}}) \quad (7)$$

through calculating the inner product of the topic embedding vector and the embedding of each word in the vocabulary. The operator $(\cdot)^T$ indicates the matrix transposition and softmax normalizes $|V_t|$ real numbers resulting from the inner product into a probability distribution of $|V_t|$ possible outcomes. With $\beta_{z_{d,n}}^t$ and $z_{d,n}$, the n -th word $w_{d,n}^t \in \{1, \dots, |V_t|\}$ can be drawn from:

$$w_{d,n}^t \sim \text{Cat}(\beta_{z_{d,n}}^t). \quad (8)$$

The log marginal likelihood of $w_{d,n}^t$ is represented as:

$$\log p(w_{d,n}^t | \mathbf{x}_d) = \log \sum_{z_{d,n}} [p(w_{d,n}^t | \beta_{z_{d,n}}^t) p(z_{d,n} | \theta_d)] = \log \mathcal{B}^t \theta_d, \quad (9)$$

where $\mathcal{B}^t = [\beta_1^t, \dots, \beta_k^t, \dots, \beta_K^t]$ is the modality t 's word probability distributions for all topics (Miao et al., 2017).

3.3 Optimization objective

We adopt the framework of variational inference (Blei et al., 2017) so that the evidence lower bound (ELBO) (Yang, 2017) for minimization is defined as:

$$\mathcal{L}_d = -\mathbb{E}_{q(\mathbf{x}_d | \mathbf{w}_d^*)} \left[\sum_{t=1}^4 \sum_{n=1}^{N_d^t} \log p(w_{d,n}^t | \mathbf{x}_d) \right] + D_{KL}[q(\mathbf{x}_d | \mathbf{w}_d^*) || p(\mathbf{x}_d)], \quad (10)$$

where N_d^t is the total number of words in modality t from the d -th EHR, and $D_{KL}[\cdot]$ measures the KL divergence (Joyce, 2011) between the prior $p(\mathbf{x}_d)$ and the variational approximation $q(\mathbf{x}_d | \mathbf{w}_d^*)$. Here, the prior $p(\mathbf{x}_d)$ is assumed to follow the normal distribution $N(0, I)$ (Miao et al., 2016).

It is worth noting that due to the high complexity of EHRs, it is not a trivial task to generalize our model to unseen EHR data. Therefore, a dropout-based regularization mechanism is introduced as a training trick to improve the representation ability. \mathbf{h}_d^{mix} is fed into two parallel sub-modules of the same architecture (composed of MLP modules with dropout) but with distinct parameters. The outputs of these two sub-modules follow two distributions $q(\mathbf{x}_d^1 | \mathbf{w}_d^*) = N(\mu_d^1, (\sigma_d^1)^2)$ and $q(\mathbf{x}_d^2 | \mathbf{w}_d^*) = N(\mu_d^2, (\sigma_d^2)^2)$. The samples are generated as follows:

$$\hat{\mathbf{x}}_d^1 = \mu_d^1 + \epsilon_1 \cdot \sigma_d^1, \quad (11)$$

and

$$\hat{\mathbf{x}}_d^2 = \mu_d^2 + \epsilon_2 \cdot \sigma_d^2, \quad (12)$$

where ϵ_1 and ϵ_2 are small random noises following normal distribution $\mathcal{N}(0, I)$. Then the average of \hat{x}_d^1 and \hat{x}_d^2 gives \hat{x}_d , that is:

$$\hat{x}_d = \frac{1}{2}(\hat{x}_d^1 + \hat{x}_d^2). \quad (13)$$

Therefore, with the adoption of the dropout-based regularization mechanism, \hat{x}_d will be obtained from Eq. (13) instead of Eq. (3). The bidirectional KL divergence from the outputs of the dropout-based regularization module will also be included in our optimization objective as:

$$\mathcal{L} = \mathcal{L}_d + \lambda_{kl} * \mathcal{L}_{KL} + \lambda_t * \mathcal{L}_{tcl} + \lambda_m * \mathcal{L}_{mcl}, \quad (14)$$

where

$$\mathcal{L}_{KL} = \frac{1}{2} D_{KL}[q(\mathbf{x}_d^1 | \mathbf{w}_d^*) || q(\mathbf{x}_d^2 | \mathbf{w}_d^*)] + \frac{1}{2} D_{KL}[q(\mathbf{x}_d^2 | \mathbf{w}_d^*) || q(\mathbf{x}_d^1 | \mathbf{w}_d^*)]. \quad (15)$$

λ_{kl} , λ_t and λ_m are the weights to balance various losses. We select their values from $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ using the grid search method. The whole training process of ConMEHR is shown in Algorithm 1.

Algorithm 1 The training process of ConMEHR

- 1: Initialize model parameters.
 - 2: Get the embedding matrix of modality t 's vocabulary, ρ^t , by text encoder.
 - 3: **for** iteration $i = 1, 2, \dots$ **do**
 - 4: Select a minibatch B of documents.
 - 5: **for** the d -th EHR in B **do**
 - 6: Get the normalized BoW representations of four modalities,
 - 7: $\mathbf{w}_d^* = \{\mathbf{w}_d^N, \mathbf{w}_d^D, \mathbf{w}_d^S, \mathbf{w}_d^M\}$.
 - 8: Get latent representation $\mathbf{h}_d^N, \mathbf{h}_d^D, \mathbf{h}_d^S$, and \mathbf{h}_d^M from \mathbf{w}_d^* through
 - 9: separated MLP modules.
 - 10: Concatenate $\mathbf{h}_d^N, \mathbf{h}_d^D, \mathbf{h}_d^S$, and \mathbf{h}_d^M to get \mathbf{h}_d^{mix} as defined in Eq. (2).
 - 11: Compute $\mu_d^1, \sigma_d^1, \mu_d^2$, and σ_d^2 through dropout-based regularization
 - 12: module.
 - 13: Sample \hat{x}_d^1 and \hat{x}_d^2 from $N(\mu_d^1, (\sigma_d^1)^2)$ and $N(\mu_d^2, (\sigma_d^2)^2)$ using
 - 14: Eq. (11–12).
 - 15: Compute $\hat{x}_d = \frac{1}{2}(\hat{x}_d^1 + \hat{x}_d^2)$ as shown in Eq. (13).
 - 16: Compute the topic distribution $\theta_d = \text{softmax}(\hat{x}_d)$ as defined in
 - 17: Eq. (4).
 - 18: **for** the n -th word in modality t **do**
 - 19: Get the topic assignment $z_{d,n} \sim \text{Cat}(\theta_d)$ from Eq. (6).
 - 20: Compute the word probability distribution for topic $z_{d,n}$ as
 - 21: $\beta_{z_{d,n}}^t = \text{softmax}((\rho^t)^T \alpha_{z_{d,n}})$ from Eq. (7).
 - 22: Compute the log marginal likelihood as $\log p(\mathbf{w}_{d,n}^t | \mathbf{x}_d) =$
 - 23: $\log \mathcal{B}^t \theta_d$ from Eq. (9), where $\mathcal{B}^t = [\beta_1^t, \dots, \beta_k^t, \dots, \beta_K^t]$.
 - 24: **end for**
 - 25: **end for**
 - 26: Train our model using the optimization objective defined in Eq. (14).
 - 27: Update model parameters by minimizing the loss function.
 - 28: **end for**
-

4 Experiments

This section conducts extensive experiments to demonstrate the performance of our proposed model on patient stratification. In Sect. 4.1, we introduce two real-world datasets, comparative methods, and evaluation metrics. In Sect. 4.2, the optimal values of hyperparameters for each comparative method are determined. In Sect. 4.3, we show quantitative evaluation results to demonstrate the superiority of our ConMEHR model on both datasets. In Sect. 4.4, we further illustrate the impacts of our method in decision support through qualitative experiments. In Sect. 4.5, we conduct ablation studies to understand the effectiveness of core modules in our model.

4.1 Experience setup

4.1.1 Dataset

To demonstrate the applicability of our proposed model in real-world scenarios, we use two EHR datasets: the MIMIC-III dataset⁷ and the Chinese Medical Clinical (CMC) dataset. We focus on four representative modalities to learn multimodal topics from EHRs: Note, Disease, Symptom, and Medication. These modalities can be divided into two categories: unstructured texts and structured medical terms. Note, which contains unstructured texts, belongs to the unstructured modality, while Disease, Symptom, and Medication consist of standard terms and are considered structured modalities. Unstructured texts can be challenging to work with because of the presence of synonyms, abbreviations, and informal language, which often make the text more difficult to process. To preprocess the unstructured texts in EHRs, we use standard NLP techniques such as word segmentation, stop word removal, and abbreviation replacement.

The details of these two datasets are summarized in Table 2. In the MIMIC-III dataset, the disease modality contains ICD codes (Organization et al., 1978), while the symptom modality is composed of Human Phenotype Ontology (HPO) codes (Köhler et al., 2017). The explanations of ICDs and HPOs are fed into the pretrained language model BioBERT (Lee et al., 2020) to generate the corresponding embeddings of these codes. BioBERT, as a domain-specific language model, is also used to embed the Note and Medication in MIMIC-III. As for the CMC dataset, it collects EHRs from Chinese medicine clinics in Hong Kong,⁸ and all information stored in CMC is in Chinese, motivating us to adopt ChineseBERT (Cui et al., 2021) for embedding.

4.1.2 Comparative methods

First, we assess the performance of ConMEHR by comparing it with the following topic modelling techniques.

- **LDA** (Blei, 2012) is a conventional topic modelling approach that assumes each document is generated from a small number of latent topics, and each topic is explained by a distribution over words. We use the implementation in gensim,⁹ a Python library.

⁷ <https://physionet.org/content/mimiciii/1.4/>

⁸ Patient consents were obtained. As an observational study, patients were not impacted by our research. All sensitive patient information was removed, and all patients are de-identified.

⁹ <https://pypi.org/project/gensim/>

Table 2 The summary of the two datasets

Dataset	MIMIC-III	CMC
# of EHRs	13492	6868
# of patients	11400	956
Vocabulary size of note	8437	4657
Vocabulary size of disease	565	607
Vocabulary size of symptom	559	387
Vocabulary size of medication	2448	622

- **NVDM** (Miao et al., 2016) is a neural variational document model that employs a continuous stochastic document representation in combination with a generative model based on a variational auto-encoder. We use the PyTorch implementation¹⁰ for comparison.
- **ProLDA** [31] is a topic modelling approach that utilizes the logistic normal distribution as the prior of topic proportion and adopts amortized variational inference. We use the original implementation¹¹ for comparison.
- **ETM** (Dieng et al., 2020) is a document generative model that combines topic models with word embeddings. We replace the Word2Vec module in ETM with BioBERT (Lee et al., 2020) and ChineseBERT (Cui et al., 2021), which have been pre-trained on large corpora, to obtain more accurate semantic embeddings for MIMIC-III and CMC data, respectively. We use the implementation in.¹²
- **ETM-multi** is an extended version of ETM that has multiple parallel data encoding and reconstruction branches for different modalities. The latent representations of different modalities are integrated together to obtain the topic distribution. Each modality is reconstructed from the topic distribution via its own reconstruction module.
- **MCSE-TM** integrates a contrastive multimodal objective, derived from the state-of-the-art MCSE model (Zhang et al., 2022), into ConMEHR's topic modelling framework. This objective allows the integration of data from multiple modalities into a shared embedding space, thereby enhancing the overall performance of the model.

Then, we conduct ablation studies to investigate the impacts of the CL modules in our ConMEHR by removing CL modules:

- **ConMER⁻** is the ablated version of ConMEHR in which both the topic-level and modality-level CL modules are removed.
- **ConMEHR⁻+ModalityCL** keeps the modality-level CL module of ConMEHR while the topic-level CL module is discarded.
- **ConMEHR⁻+TopicCL** keeps the topic-level CL module of ConMEHR while the modality-level CL module is discarded.

4.1.3 Evaluation metrics

The performance of all comparative models is evaluated using the following four metrics:

- **Coherence.** Topic coherence measures the degree of similarity between the top words within each topic. It is calculated by averaging the pointwise mutual information of pairs

¹⁰ <https://github.com/YongfeiYan/Neural-Document-Modeling>

¹¹ <https://github.com/akashgit/autoencodingvifortopicmodels>

¹² <https://github.com/adjidieng/ETM>

Table 3 Two illustrative examples from the MIMIC-III dataset

Modality	EHR	
	EHR1	EHR2
Note	‘You were admitted to the hospital for an enlarging infection. It is not amenable to surgery...’	‘You were admitted to the hospital after feeling weak and having fevers for several days. You had developed new kidney failure...’
Disease	ICD4589: ‘Hypotension, unspecified’, ICD5849: ‘Acute kidney failure, unspecified’ ,...	ICD5856: ‘End stage renal disease’, ICD5849: ‘Acute kidney failure, unspecified’ ,...
Symptom	HP:0001919: ‘sudden loss of renal function...’ , HP:0100750: ‘collapse of part of a lung associated with absence of inflation (air) of that part.’,...	HP:0003774: ‘a degree of kidney failure severe enough to require dialysis or kidney transplantation for survival...’ ,...
Medication	‘Sodium Chloride’, ‘Dextrose’, ‘Ciprofloxacin HCl’,...	‘Ciprofloxacin HCl’, ‘Meperidine’, ‘Sodium Chloride’,...

of words (Dieng et al., 2020). Topic coherence is a measure of how closely related the top words in a given topic are, and it is typically quantified by computing the average pointwise mutual information between pairs of words within that topic. Essentially, this means that the higher the coherence score for a given topic, the more semantically similar and connected its top words are to each other.

As mutual information measures the occurrence of a word pair, it emphasizes their semantic similarity. Table 3 provides an illustrative example. ‘EHR1’ and ‘EHR2’ are two samples in MIMIC-III containing the disease code ‘ICD5849’. The HPO code ‘HP:0001919’ from ‘EHR1’ and ‘HP:0003774’ from ‘EHR2’ have similar semantic meanings. When calculating the mutual information between ‘ICD5849’ and ‘HP:0001919’, only their co-occurred EHRs will be considered so that samples like ‘EHR2’ will be excluded.

Hence, we adopt a new measure of coherence, which replaces mutual information with cosine similarity. Specifically, the *Coherence* score is calculated as:

$$Coherence = \frac{1}{K} \sum_{k=1}^K \frac{1}{45} \sum_{i=1}^{10} \sum_{j=i+1}^{10} g_c(\mathbf{w}_i^{(k)}, \mathbf{w}_j^{(k)}), \quad (16)$$

where $\{\mathbf{w}_1^{(k)}, \dots, \mathbf{w}_{10}^{(k)}\}$ denotes the latent representations of the top 10 most likely words in the topic k , and $g_c(\cdot, \cdot)$ calculates the cosine similarity between two words. A large *Coherence* value indicates that the top 10 words of the same topic are quite likely to be semantically similar.

- **Diversity.** It measures the degree of diversity among various latent topics. Following the definition in Dieng et al. (2020), it is calculated as the percentage of unique words in the top words of all topics. If the *Diversity* score approaches 1.0, then we can say that the learned topics have no significant overlaps.

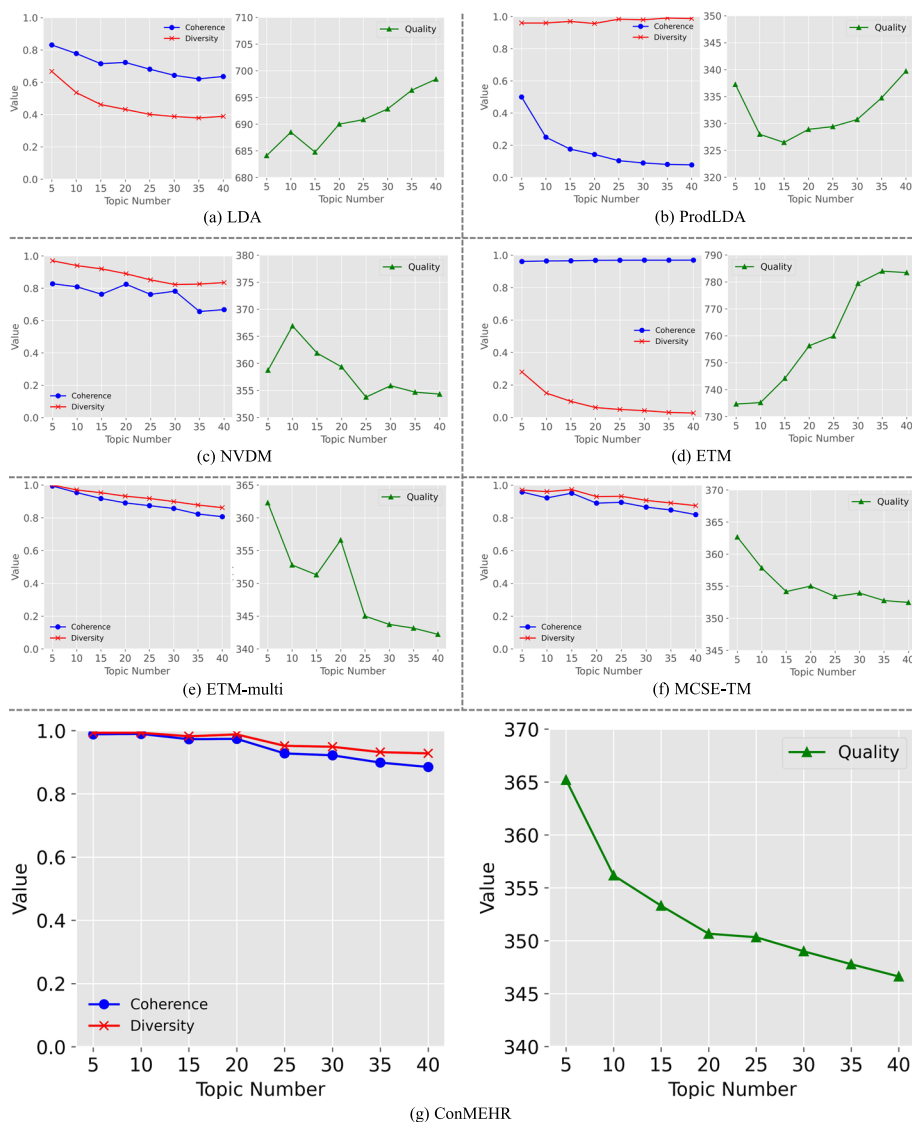


Fig. 5 The values of *Coherence*, *Diversity* and *Quality* in the MIMIC-III dataset with different values of topic number K

- *Quality*. It shows the quality of the BoW reconstruction of the topic model. For each document in the dataset, we obtain the document-to-topic distribution and use it to reconstruct the document. The *Quality* score is calculated as the average reconstruction loss of the dataset, as described in Eq. (9), following the definition proposed in Miao et al. (2016). Therefore, a lower *Quality* score means that the model can generate better document-to-topic distributions for document reconstruction.
- *Ratio*. It reflects the relative contributions from the four modalities for learning latent consensus topics and is in the form of $r_N/r_D/r_S/r_M$. To obtain the *Ratio*, we first select

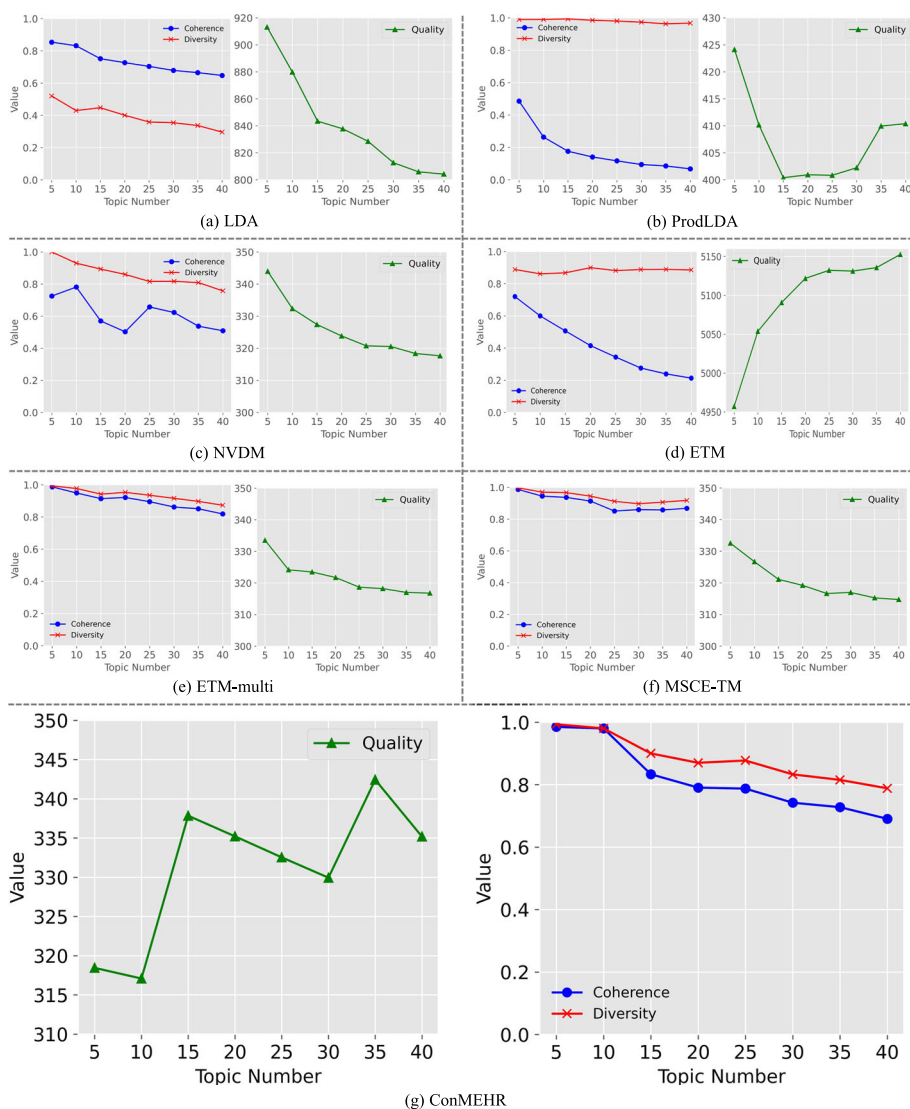


Fig. 6 The values of *Coherence*, *Diversity* and *Quality* in the CMC dataset with different values of topic number K

the 10 top-scoring words for each topic and then calculate the proportion of these $K \times 10$ top words from different modalities. For example, r_N is the proportion of top words from Note modality. By examining the *Ratio* result, we can determine whether any modalities have dominated the topic learning process while some other modalities' information is ignored due to imbalanced dimensionality across different modalities.

Table 4 The selected topic number for each comparative method

Methods	Datasets	
	MIMIC-III	CMC
LDA	5	15
ProdLDA	10	15
NVDM	20	10
ETM	5	5
ETM-multi	15	20
MCSE-TM	15	15
ConMEHR	20	10

4.2 Hyperparameter selection

For all comparative models, the topic number (the number of subgroups) is a hyperparameter that needs to be predefined. Here, we investigate the influence of the topic number and select the optimal values for the following experiments. We choose the topic number from {5, 10, 15, 20, 25, 30, 35, 40}. Figures 5 and 6 show the changes in the *Coherence*, *Diversity*, and *Quality* scores for MIMIC-III and CMC, respectively, as the number of topics increases. By examining these figures, we choose the optimal topic number for each model to be the one that yields high *Coherence* and *Diversity* scores while maintaining a low *Quality* score.

Table 4 contains the optimal K for all comparative models. Let us explain the choice of our ConMEHR model as an example. As shown in Fig. 5g, when K is less than 20, the *Coherence* and *Diversity* values are close to their maximum value of 1.0 while the *Quality* value remains high. When K is greater than 20, the *Coherence* and *Diversity* values show an overall downward trend, and the *Quality* curve slows down its decrease. Therefore, we choose $K = 20$ for ConMEHR in the MIMIC-III dataset. Figure 6g shows that for the CMC dataset, the *Coherence* and *Diversity* scores decrease after the number of topics is greater than 10, while the *Quality* score reaches its optimal value at $K = 10$. Therefore, we set $K = 10$ for the CMC dataset in the following experiments. We also choose the weights of various losses in Eq. (14) using the grid search method. Finally, we selected $\lambda_{kl} = 100$, $\lambda_t = 10$, and $\lambda_m = 1$ for the MIMIC-III dataset and $\lambda_{kl} = 10$, $\lambda_t = 1$, and $\lambda_m = 1$ for the CMC dataset.

4.3 Quantitative evaluation

In this section, we employed hold-out cross-validation to ensure a fair comparison with other topic modelling methods. Although both multi-fold and hold-out cross-validation is commonly used, multi-fold is typically more suitable for small sample sizes. However, since our experimental section utilized the MIMIC-III and CMC datasets, which consist of 13,492 and 6,868 samples respectively, hold-out cross-validation was the appropriate choice for our study. Additionally, to reduce the impact of randomness, we trained all models five times with a fixed set of five different seeds and presented indicator performance and standard deviation as in Table 5.

First, we compare ConMEHR with models that only handle single-modality data, where we concatenate inputs from the four modalities into one sequence as the input. We observe that our ConMEHR model outperforms LDA, ProdLDA, NVDM, and ETM for both datasets, achieving the best *Coherence* and *Diversity* values. For the *Quality* score, ConMEHR has the

Table 5 Comparison in topic modelling on the MIMIC-III and CMC datasets. ↓ (or ↑) indicates smaller (or larger) values are preferred. The best results are highlighted in **bold**

Datasets	Models	Evaluation metrics			Ratio ($r_N/r_D/r_S/r_M$)
		Coherence ↑	Diversity ↑	Quality ↓	
MIMIC-III	LDA	0.831±0.015	0.667±0.008	684.1±2.624	0.74/0.16/0.04/0.07
	ProdLDA	0.176±0.001	0.972±0.003	326.5±3.466	0.61/0.20/0.03/0.17
	NVDM	0.824±0.033	0.828±0.028	358.7±0.262	0.38/0.20/0.04/0.38
	ETM	0.961±0.001	0.284±0.005	734.7±6.308	0.73/0.17/0.05/0.05
	ETM-multi	0.918±0.012	0.950±0.003	351.3±0.180	0.24/0.20/0.20/0.36
	MCSE-TM	0.951±0.015	0.973±0.011	354.17±0.212	0.23/0.22/0.23/0.32
	ConMEHR	0.968±0.015	0.995±0.012	350.6±0.781	0.26/0.22/0.20/0.33
	LDA	0.752±0.014	0.442±0.016	843.4±7.153	0.63/0.08/0.01/0.28
CMC	ProdLDA	0.264±0.005	0.993±0.001	410.8±0.664	0.72/0.08/0.10/0.11
	NVDM	0.537±0.037	0.812±0.016	321.4±2.902	0.64/0.06/0.09/0.22
	ETM	0.889±0.010	0.725±0.070	4957±16.42	0.54/0.05/0.03/0.39
	ETM-multi	0.921±0.014	0.953±0.012	321.7±0.212	0.19/0.20/0.27/0.34
	MCSE-TM	0.937±0.016	0.967±0.009	321.1±0.342	0.20/0.18/0.26/0.36
	ConMEHR	0.986±0.013	0.998±0.001	320.6±4.764	0.18/0.21/0.30/0.31

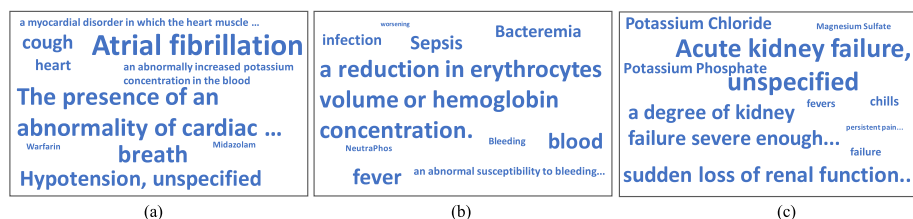


Fig. 7 The word clouds of the top words in different topics. **a** Topic T_1 ; **b** Topic T_2 ; **c** Topic T_3

second-smallest value for MIMIC-III and the smallest value for CMC. Although ProdLDA has the lowest *Quality* score for MIMIC-III, it has a highly imbalanced *Ratio* result of 0.61/0.20/0.03/0.17, indicating that Note modality has dominated the topic learning process.

The incorporation of multiple parallel structures to handle various modalities in the ETM-multi, MCSE-TM, and ConMEHR models led to excellent performance in achieving a balanced Ratio score across both datasets when compared to LDA, ProdLDA, NVDM, and ETM. These models used parallel branches to model data from different modalities separately, enabling the effective capture of data complexity and diversity, resulting in overall improved performance.

We further investigated multi-modal models with different CL strategies. When comparing MCSE-TM with ETM-multi, we observed that MCSE-TM, which integrates a contrastive multimodal objective that is absent in ETM-multi, exhibited superior performance in terms of *Coherence*, *Diversity*, and *Quality* metrics. The addition of a CL module facilitated the alignment of information from diverse modalities in the representation space, resulting in an overall enhancement of the topic model's performance. Although both MCSE-TM and ConMEHR incorporate CL modules, our evaluation metrics revealed that ConMEHR outperformed MCSE-TM. This indicates that our proposed CL module, which operates at both the modality-level and topic-level, is more effective than existing CL modules. The respective roles of the modality-level and topic-level in topic modelling were further explored in our ablation study, as detailed in Sect. 4.5.

4.4 Qualitative evaluation

4.4.1 Distributions of words across different topics

In order to interpret the meaning of the learned topics, we examine the distributions of words across different topics and identify the most influential words for each topic. We set the number of topics to 20 in the MIMIC-III dataset, as discussed in Sect. 4.2, resulting in 20 different topics generated from our model. For illustrative purposes, we select T_1 , T_2 , T_3 for further investigation.

Figure 7 displays the top words in the topics using a word cloud (Wilson, 1912). We can observe that T_1 , T_2 , and T_3 are related to heart, blood, and renal diseases, respectively. To further validate this finding, we list the topic words in Table 6 and categorize them based on the modalities they belong to. It is interesting to note that the top words within each topic are relevant. For instance, T_1 is associated with two diseases, namely, 'Atrial fibrillation' and 'Hypotension, unspecified'. Clinical studies have shown that heart disease is a significant factor that can cause hypotension (Gorelik et al., 2016). T_2 includes 'Sepsis' and 'Bacteremia', where previous research has found that sepsis is a stage that occurs after

Table 6 Top words of three most used topics from ConMEHR

Topic	Modalities		Disease	Symptom	Medication
	Note				
T1	'Breath',	'Atrial fibrillation',	'The presence of an abnormality of cardiac function...', 'a myocardial disorder in which the heart muscle is structurally and functionally abnormal...', 'an abnormally increased potassium concentration in the blood.'	'Warfarin',	
	'Heart'	'Hypotension, Unspecified', 'Cough'			
		'Sepsis',			
		'Bacteremia'			
T2	'Blood',		'A reduction in erythrocytes volume or hemoglobin concentration.', 'an abnormal susceptibility to bleeding...'	'NeutraPhos'	
	'Infection',				
	'Bleeding',				
	'Fever',				
T3	'Worsening'		'Sudden loss of renal function...', 'a degree of kidney failure severe enough failure...', 'persistent pain...',	'Potassium chloride', 'Potassium phosphate', Magnesium sulfate'	
	'Chills',	'Acute kidney failure,			
		Unspecified'			
	'Fever',				
	'Failure'				

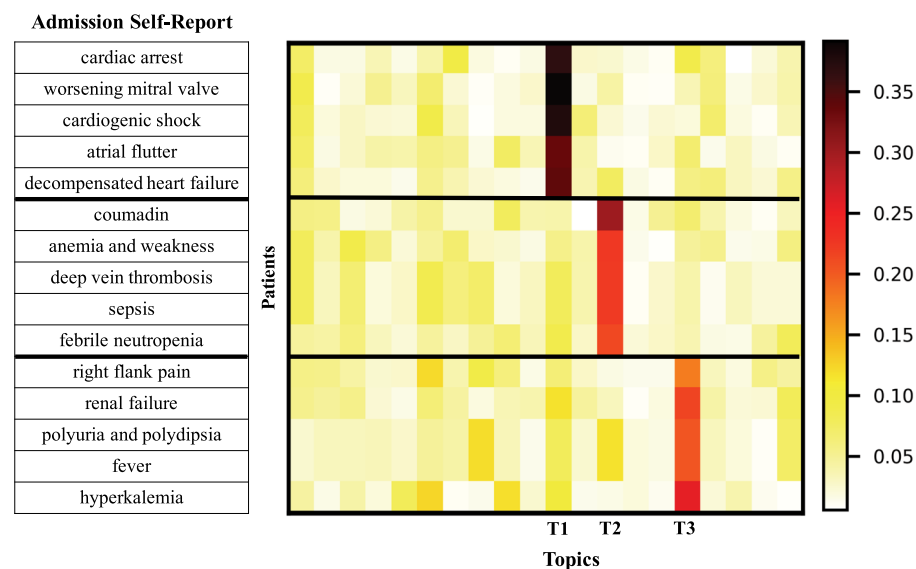


Fig. 8 The topic distributions of the top 5 patients in selected disease topics. The heatmap on the **right** indicates patient mixture memberships for the topics. The columns on the **left** show the admission self-report for each patient

the progression of bacteremia (Bone et al., 1997). Furthermore, we can observe that the top words from different modalities have connections. For example, the primary disease term in $T3$ is ‘Acute kidney failure, unspecified’, and the medication ‘Potassium Chloride’ in $T3$ can protect kidney function in patients with kidney disease (Saxena, 1989). These observations suggest that the learned topics are meaningful and can explain latent patient groups, comorbidities, associated symptoms, and potential medications

4.4.2 The characteristics of patients

In this subsection, we further explore the characteristics of patients to see whether their information is consistent with their associated topics. For each topic from $T1$, $T2$, $T3$, we select five relevant patients based on their topic distributions. Figure 8 visualizes patients’ topic distributions, where the left column indicates the corresponding admission self-reports that appeared in patients’ raw EHR. The admission self-report is the patients’ self-description of their main conditions upon admission. For example, we can see that patients in $T1$ usually present as ‘cardiac arrest’ and ‘atrial flutter’, which is consistent with the observation from previous experiments that $T1$ is associated with heart disease.

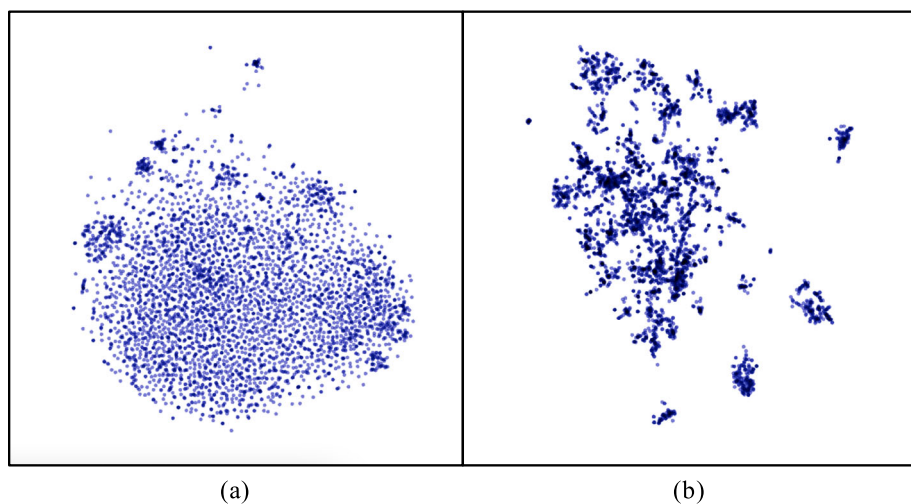
From both quantitative and qualitative results, we conclude that ConMEHR is effective in learning the meaningful subgroup from multimodal EHRs.

4.5 Ablation study

In this section, we investigate the impact of two CL modules: the topic-level CL module, which separates topic embeddings, and the modality-level CL module, which assists in multimodal data aggregation. We evaluate three ablated versions of our model: ConMEHR[−], which

Table 7 Effectiveness of CL modules. ↓ (or ↑) indicates smaller (or larger) values are preferred

Datasets	Models	Evaluation Metrics	
		Coherence ↑	Diversity ↑
MIMIC-III	ConMEHR	0.968±0.015	0.995±0.012
	ConMEHR [−]	0.919±0.018	0.951±0.001
	ConMEHR [−] +ModalityCL	0.953±0.011	0.972±0.010
	ConMEHR [−] +TopicCL	0.962±0.010	0.978±0.014
CMC	ConMEHR	0.986±0.013	0.998±0.001
	ConMEHR [−]	0.963±0.012	0.982±0.008
	ConMEHR [−] +ModalityCL	0.982±0.011	0.992±0.008
	ConMEHR [−] +TopicCL	0.978±0.009	0.986±0.009

**Fig. 9** The t-SNE visualization of patient embeddings of ConMEHR[−] and ConMEHR models in the MIMIC-III datasets. **a** patient embeddings of ConMEHR[−] in the MIMIC-III dataset; **b** patient embeddings of ConMEHR in the MIMIC-III dataset

removes both CL modules; ConMEHR[−]+ModalityCL, which removes the topic-level CL module; and ConMEHR[−]+TopicCL, which removes the modality-level CL module. Table 7 presents the *Coherence* and *Diversity* scores of all ablated models for performance comparison. We observe that the full model achieves the best performance, while the ablated version without any CL modules has the lowest *Coherence* and *Diversity* scores. Removing either the modality-level or the topic-level CL module leads to a deterioration in performance.

In addition to quantitative evaluation, we also use visualizations to examine whether CL helps in learning separable representations of patients. We use the popular t-SNE method to project high-dimensional representations into low-dimensional vectors (Van der Maaten & Hinton, 2008). Figure 9 displays the t-SNE visualization results for the MIMIC-III dataset, where we adopt default settings. Figure 9a and b are the results from ConMEHR[−] and ConMEHR, respectively. It is evident that ConMEHR generates more separable patient representations, demonstrating the effectiveness of CL.

5 Discussion and conclusions

We develop the ConMEHR model to explain patient subgroups by extracting meaningful topics from multimodal EHRs. In ConMEHR, the modality-level and topic-level CL modules are adopted to obtain a unified latent space for multiple modalities and diversify patient subgroups, respectively. We compare the performance of ConMEHR with several topic modelling methods on two real-world EHR datasets and demonstrate that ConMEHR performs well in generating topics. The key aspects of our study are as follows.

Technical Status: To illustrate the technical status of our patient stratification approach, we highlight its differences from classification models such as Bagging CART and XGBoost-based methods (Affes & Hentati-Kaffel, 2019; Du Jardin, 2021; du Jardin, 2022) commonly used in decision support. Specifically, we emphasize two key aspects. On the one hand, patient stratification is the process of classifying patients into different groups based on their shared medical conditions. However, unlike decision tree methods such as Bagging CART and XGBoost, which primarily focus on predicting patients into pre-defined classes, our patient stratification approach goes beyond classification by explaining each subgroup based on common medical conditions among patients within the same group. This can help clinicians and researchers understand the underlying factors contributing to a specific patient's condition and provide relevant prognoses. On the other hand, unlike our patient stratification algorithm, Bagging CART and XGBoost cannot incorporate semantic representation in their decision-making process. This means that our model considers the meaning and context of the data, as well as any relevant background knowledge, to generate more accurate and meaningful patient groupings. By incorporating semantic representation, it can also identify underlying patterns and correlations that may be missed by traditional statistical methods, leading to more precise and effective patient stratification.

Technical Contribution: There are two highlights in this research. The first one is the modality-level CL module in ConMEHR, which can learn a unified latent representation space for multiple modalities and integrate latent representations to obtain patient subgroups. The second highlight is the topic-level CL module, which guarantees diverse topic representations.

Practical Implementation: Our approach provides a powerful solution for patient stratification by enabling the identification of subgroups based on their medical conditions. The model does not require re-training unless new patients have medical conditions that are not already present in the training data. Considering the existence of a vast EHR corpus like 13,492 EHRs in MIMIC-III, it is highly likely that new patients' medical conditions or terms will already be present within the corpus. Leveraging the large EHR corpus is key to enabling our approach to scale effectively and generate accurate patient classifications. In the uncommon event that new patients present with medical conditions that were not included in the training data, our approach offers a practical solution. Rather than starting from scratch, the existing model can be fine-tuned to incorporate the new conditions, reducing the need for a complete re-training of the model. This fine-tuning process leverages the existing knowledge and expertise captured in the model, enabling it to adapt to new scenarios more efficiently. This feature provides a highly adaptable and flexible solution for patient stratification, ensuring that the model can accommodate any new medical conditions that arise over time.

Applicable Significance: Our work can automatically detect subgroups of patients, which is an important task in the domain of precision medicine. This will have a direct impact on patients, healthcare providers, and researchers. Specifically, it would significantly accelerate the decision process and help to design treatments tailored to each subgroup. Such an impact will last since the research outcome as a new modelling approach could help achieve precision

medicine by leveraging information from EHRs. Thus, it has the potential to be broadly applied in future healthcare.

Future Direction: In the future, we will also extend our model to incorporate more modalities and also medical knowledge graphs. We will develop a concrete system and provide it to clinicians. With our patient stratification system, clinicians would further accelerate their decision process. Patients will be, for example, diagnosed with diseases at early stages and assessed for disease subtypes and prospective risks. Meanwhile, we will develop models using not only EHR data but also other patients' health-related data collected in daily life. By learning the trajectories of patients and monitoring their changes in subgroups, we can send alerts to patients and clinicians when some health states are changed. In this way, we believe our work would help to boost the development of precision medicine.

Acknowledgements This work is supported by the National Key Research and Development Program of China (No. 2021ZD0113303), and the National Natural Science Foundation of China (Nos. 62022052).

Declarations

competing interests The authors have no competing interests to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Affes, Z., & Hentati-Kaffel, R. (2019). Forecast bankruptcy using a blend of clustering and mars model: Case of us banks. *Annals of Operations Research*, 281(1–2), 27–64.
- Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., & Zhu, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *International conference on machine learning* (pp. 280–288). PMLR
- Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., & Saunshi, N. (2019). A theoretical analysis of contrastive unsupervised representation learning. In *36th International conference on machine learning, ICML 2019* (pp. 9904–9923). International Machine Learning Society (IMLS).
- Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. In *Advances in neural information processing systems*, vol. 32.
- Benson, A. R., Lee, J. D., Rajwa, B., & Gleich, D. F. (2014). Scalable methods for nonnegative matrix factorizations of near-separable tall-and-skinny matrices. In *Advances in neural information processing systems*, vol. 27.
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.
- Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518), 859–877.
- Bone, R. C., Grodzin, C. J., & Balk, R. A. (1997). Sepsis: A new hypothesis for pathogenesis of the disease process. *Chest*, 112(1), 235–243.
- Bunk, S., & Krestel, R. (2018). Welda: Enhancing topic models by incorporating local word context. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries* (pp. 293–302).
- Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33, 9912–9924.


- Chechik, G., Sharma, V., Shalit, U., & Bengio, S. (2010). Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3), 1109–1135.
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607). PMLR.
- Chi, E. C., & Kolda, T. G. (2012). On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4), 1272–1299.
- Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (vol. 1, pp. 539–546). IEEE.
- Cong, Y., Chen, B., Liu, H., & Zhou, M. (2017). Deep latent Dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC. In *International conference on machine learning* (pp. 864–873). PMLR.
- Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-training with whole word masking for Chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3504–3514.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (vol. 1, pp. 4171–4186). Long and Short Papers.
- Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8, 439–453.
- Ding, W., Ishwar, P., & Saligrama, V. (2015). Most large topic models are approximately separable. In *2015 Information theory and applications workshop (ITA)* (pp. 199–203). IEEE.
- Du Jardin, P. (2021). Forecasting bankruptcy using biclustering and neural network-based ensembles. *Annals of Operations Research*, 299(1–2), 531–566.
- du Jardin, P. (2022). Designing topological data to forecast bankruptcy using convolutional neural networks. *Annals of Operations Research*, 325, 1–42.
- Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., & Goldberg, Y. (2021). Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9, 1012–1031.
- Gao, T., Yao, X., & Chen, D. (2021). SIMCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 6894–6910).
- Gorelik, O., Feldman, L., & Cohen, N. (2016). Heart failure and orthostatic hypotension. *Heart Failure Reviews*, 21(5), 529–538.
- Grill, J.-B., Strub, F., Althé, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent—a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33, 21271–21284.
- Harshman, R. A. (1972). Note: This manuscript was originally published in 1972 and is reproduced here to make it more accessible to interested scholars. The original reference is Harshman, RA (1972). In *PARAFAC2: Mathematical and technical notes. UCLA working papers in phonetics* (pp. 30–44). University Microfilms, Ann Arbor, Michigan.
- He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9729–9738).
- Henaff, O. (2020). Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning* (pp. 4182–4192). PMLR.
- Henderson, J., Ho, J. C., Kho, A. N., Denny, J. C., Malin, B. A., Sun, J., & Ghosh, J. (2017). Granite: Diversified, sparse tensor factorization for electronic health record-based phenotyping. In *2017 IEEE International conference on healthcare informatics (ICHI)* (pp. 214–223). IEEE.
- Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., & Bengio, Y. (2018). Learning deep representations by mutual information estimation and maximization. In *International conference on learning representations*.
- Hoffer, E., & Ailon, N. (2015). Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition* (pp. 84–92). Springer.
- Ho, J. C., Ghosh, J., Steinhubl, S. R., Stewart, W. F., Denny, J. C., Malin, B. A., & Sun, J. (2014). Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics*, 52, 199–211.
- Ho, C.-H., & Nvasconcelos, N. (2020). Contrastive learning with adversarial examples. *Advances in Neural Information Processing Systems*, 33, 17081–17093.

- Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1), 1–9.
- Joyce, J. M. (2011). Kullback-leibler divergence. In *International encyclopedia of statistical science* (pp. 720–722). Springer.
- Kenton, J. D. M., -W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186).
- Kim, Y., Sun, J., Yu, H., & Jiang, X. (2017). Federated tensor factorization for computational phenotyping. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 887–895).
- Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint [arXiv:1312.6114](https://arxiv.org/abs/1312.6114)
- Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S. M., Boerkoel, C. F., Boycott, K. M., et al. (2017). The human phenotype ontology in 2017. *Nucleic Acids Research*, 45(D1), 865–876.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240.
- Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., & Wang, H. (2021). Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In: *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (vol. 1, pp. 2592–2607). Long Papers.
- Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21(3), 105–117.
- Logeswaran, L., & Lee, H. (2018). An efficient framework for learning sentence representations. In *International conference on learning representations*.
- Miao, Y., Grefenstette, E., & Blunsom, P. (2017). Discovering discrete latent topics with neural variational inference. In *International conference on machine learning* (pp. 2410–2419). PMLR.
- Miao, Y., Yu, L., & Blunsom, P. (2016). Neural variational inference for text processing. In *International conference on machine learning* (pp. 1727–1736). PMLR.
- Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space.
- Nwankpa, C. E., Ijomah, W., Gachagan, A., & Marshall, S. (2021). Activation functions: comparison of trends in practice and research for deep learning. In *2nd International conference on computational sciences and technology*.
- Oh Song, H., Xiang, Y., Jegelka, S., & Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4004–4012).
- Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint [arXiv:1807.03748](https://arxiv.org/abs/1807.03748)
- Organization, W. H., et al. (1978). International classification of diseases: [9th] revision, basic tabulation list with alphabetic index. In *International classification of diseases: 9th revision, basic tabulation list with alphabetic index*.
- Pan, T., Song, Y., Yang, T., Jiang, W., & Liu, W. (2021). Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11205–11214).
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982–3992).
- Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In: *International conference on machine learning* (pp. 1278–1286). PMLR.
- Saxena, K. (1989). Clinical features and management of poisoning due to potassium chloride. *Medical Toxicology and Adverse Drug Experience*, 4(6), 429–443.
- Shi, B., Lam, W., Jameel, S., Schockaert, S., & Lai, K. P. (2017). Jointly learning word embeddings and latent topics. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (pp. 375–384).
- Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., & Kiela, D. (2022). Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15638–15650).

- Srivastava, A., & Sutton, C. (2017). Autoencoding variational inference for topic models. In *International conference on learning representations*.
- Tian, Y., Krishnan, D., & Isola, P. (2020). Contrastive multiview coding. In *European conference on computer vision* (pp. 776–794). Springer.
- Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2579–2605.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).
- Wang, Y., Chen, R., Ghosh, J., Denny, J. C., Kho, A., Chen, Y., Malin, B. A., & Sun, J. (2015). Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1265–12740).
- Wang, Y., Benavides, R., Diatchenko, L., Grant, A. V., & Li, Y. (2022). A graph-embedded topic model enables characterization of diverse pain phenotypes among UK biobank individuals. *iScience*, 25, 104390.
- Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(2), 207–244.
- Wilson, S. (1912). Word cloud contest. *Brain*, 34(1), 295–507.
- Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3733–3742).
- Wu, L., Li, J., Wang, Y., Meng, Q., Qin, T., Chen, W., Zhang, M., Liu, T.-Y., et al. (2021). R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34, 10890–10905.
- Xu, Z., So, D. R., & Dai, A. M. (2021). Mufasa: Multimodal fusion architecture search for electronic health records. In *Proceedings of the AAAI Conference on Artificial Intelligence* (vol. 35, pp. 10532–10540).
- Xun, G., Li, Y., Zhao, W. X., Gao, J., & Zhang, A. (2017). A correlated topic model using word embeddings. In *IJCAI* (pp. 4207–4213).
- Yang, X. (2017). Understanding the variational lower bound. *Variational Lower Bound ELBO, Hard Attention*, 13, 1–4.
- You, C., Chen, N., & Zou, Y. (2021). Self-supervised contrastive cross-modality representation learning for spoken question answering. In *Findings of the association for computational linguistics: EMNLP 2021* (pp. 28–39).
- Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). Coca: Contrastive captioners are image-text foundation models. arXiv preprint [arXiv:2205.01917](https://arxiv.org/abs/2205.01917)
- Zhang, H., Chen, B., Guo, D., & Zhou, M. (2018). Whai: Weibull hybrid autoencoding inference for deep topic modeling. In *International conference on learning representations*.
- Zhang, M., Mosbach, M., Adelani, D., Hedderich, M., & Klakow, D. (2022). Mcse: Multimodal contrastive learning of sentence embeddings. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 5959–5969).
- Zhang, D., Nan, F., Wei, X., Li, S. -W., Zhu, H., Mckeown, K., Nallapati, R., Arnold, A. O., & Xiang, B. (2021). Supporting clustering with contrastive learning. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 5419–5430).
- Zhao, H., Du, L., Buntine, W., & Liu, G. (2017). Metalda: A topic model that efficiently incorporates meta information. In *2017 IEEE international conference on data mining (ICDM)* (pp. 635–644). IEEE.
- Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., & Buntine, W. (2021). Topic modelling meets deep neural networks: A survey. In *International Joint Conference on Artificial Intelligence 2021* (pp. 4713–4720). Association for the Advancement of Artificial Intelligence (AAAI).
- Zou, Y., Pesaraghader, A., Song, Z., Verma, A., Buckeridge, D. L., & Li, Y. (2022). Modeling electronic health record data using an end-to-end knowledge-graph-informed topic model. *Scientific Reports*, 12(1), 1–14.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Qing Yin¹ · Linda Zhong² · Yunya Song³ · Liang Bai⁴ · Zhihua Wang⁵ · Chen Li⁶ · Yida Xu⁷ · Xian Yang¹ 

Qing Yin
qing.yin-2@postgrad.manchester.ac.uk

Linda Zhong
linda.zhong@ntu.edu.sg

Yunya Song
yunyasong@hkbu.edu.hk

Liang Bai
bailiang@sxu.edu.cn

Zhihua Wang
zhihua.wang@zju.edu.cn

Chen Li
lichen@hust.edu.cn

Yida Xu
xuyida@hkbu.edu.hk

- ¹ Alliance Manchester Business School, University of Manchester, Oxford Road, Manchester M139PL, United Kingdom
- ² School of Biological Sciences, Nanyang Technological University, Singapore, Singapore
- ³ AI and Media Research Lab, Hong Kong Baptist University, Hong Kong, China
- ⁴ School of Computer and Information Technology, Shanxi University, Taiyuan, China
- ⁵ China Shanghai Institute for Advanced Study of Zhejiang University, Zhejiang, China
- ⁶ School of Physical Education, Huazhong University of Science and Technology, Wuhan, China
- ⁷ Department of Mathematics, Hong Kong Baptist University, Hong Kong, China